Sead Harmandić

Analysis and Visualization of Real-Time Twitter Data

Master's Thesis

Graz University of Technology

Institute for Information Systems and Computer Media
Head: Prof. PhD Frank Kappe

Supervisor: Assoc. Prof. PhD Martin Ebner

Graz, August 2015

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, ______________________ ______________________

Date Signature

Eidesstattliche Erklärung[1]

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am ______________________ ______________________

Datum Unterschrift

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Acknowledgements

First and foremost, I would like to thank my supervisor Assoc. Prof. PhD Martin Ebner for all the useful comments, guidelines and his engagement throughout the research process as well as for giving me an opportunity to join his development team during my studies.

Furthermore, I would like to express my gratitude to everyone who has supported me at all levels of the research project.

And last but not least, I would like to express my gratitude to my parents Alija and Mubera and my brother Adnan for their endless support and encouragement.

Thank you,

Sead Harmandić

Abstract

The use of social media, especially Twitter, for real-time communication is growing each day. In order to extract the most of the Twitter data, one must provide appropriate data mining of the tweets. Proper and meaningful analysis of the tweets with the visual representation of the results is only possible if the tweets can be collected and analyzed. The process of data manipulation must include both old and new tweets. The visualization of the tweets of a particular event is the main objective of this paper and will be explained in detail.

The research objectives of this paper provide an overview of the currently existing scientific papers as well as other available tools for data analysis and visualization with a special focus on Twitter. The variety of the usage of tweets is presented in the aforementioned scientific papers and possible application fields have also been proposed.

Based on the existing tools and their functionality, a new tool called TwitterSuitcase was developed for the needs of tweets analysis and visualization in this master's thesis. An application called TweetCollector which ensures specific tweet collections served as basis software for the development of this tool.

The evaluation of TwitterSuitcase is performed by processing real-time use case, describing each of the created results separately and discussing attained results. The future of Twitter and its data is also discussed in terms of scientific results. The final conclusion of the master's thesis leads to the conclusion that social media will continue providing important information and that the tools will have to be more efficient in order to extract the maximum out of the shared information.

Kurzfassung

Der Einsatz von Social Media, vor allem Twitter für die Echtzeitkommunikation wächst mit jedem Tag. Um das Meiste aus den Twitter-Daten herauszuholen werden adäquate Data-Mining Prozesse benötigt. Eine effiziente und geeignete Analyse der Tweets ist nur dann möglich wenn die Tweets gesammelt, analysiert und dargestellt werden können. Der Prozess der Datenmanipulation muss sowohl alte als auch neue Twitter-Daten involvieren. Die Visualisierung der Tweets gehört daher zu den Hauptaufgabe dieser Arbeit und wird detailliert erklärt.

Dazu werden entsprechend vorhandene Werkzeuge zur Datenanalyse und Datenvisualisierung analyisiert und auf ihre Tauglichkeit verglichen. Verschiedene wissenschaftliche Artikel werden als Basis für die Vielzahl der Twitter-Anwendungen und möglichen Einsatzbereiche herangezogen.

Ein neues Werkzeug genannt TwitterSuitcase wird, auf Basis bereits vorhandener Werkzeuge für die Bedürfnisse der Analyse und Visualisierung innerhalb dieser Arbeit, erstellt. Die Grundsteine dieses Werkzeugs werden, in Form von Tweet-Daten durch die Applikation TweetCollector gewährleistet.

Die Evaluierung des TwitterSuitcases wird anhand eines Anwendungsfalles dargestellt. Dabei werden die Ergebnisse einzeln präsentiert und anschließend darüber diskutiert. Die Zusammenfassung dieser Arbeit führt zum Schluss, dass Social Media weiterhin eine wichtige Rolle spielen wird. Das wird vor allem möglich sein durch die Bereitstellung wichtiger Daten und Werkzeuge um das Wesentliche und Wichtigste aus diesen Daten herauszuholen.

Contents

List of Figures

1. Introduction

Twitter is an online social networking service (also called Social Networking Site or SNS) launched in 2006. The interest in Twitter has been high from the very beginning, and the total number of active Twitter users has risen up to over 140 million by 2012 [Twitter, 2015]. Today, nine years after the service started, the number of active users is over 300 millions [Chris Welch, 2015]. It is one of the most popular micro-blogging services in the world which increased the popularity of micro-blogging platform. Twitter gives every single person a possibility to directly and publicly communicate with other persons, politicians, corporations or any other large entities. For example, the importance of such communication is also recognized by the White House, which is publishing the latest news from President Obama and his administration[1]. On the other hand, the usage of Twitter for bringing up opinions or information in real-time was demonstrated during the Arab Spring in year 2010 when it was one of the most powerful tools in the fight against the government [Yousri Marzouki, 2015]. Some of the countries, such as China, Iran and North Korea have a long history of social media censorship due to the fact that social media was often used as a platform to coordinate protests and to spread real-time information identified by governments as harmful and threatening the system [Ng, 2013]. The fact that Twitter has such a huge influence is one of the reasons why it is interesting to analyze the data further more. In the past, a number of research on this topic has been conducted [Honeycutt and Herring, 2009, Boyd et al., 2010, Ebner, 2013]. A useful and meaningful set of analyzed data includes the latest information, as well the older information related to same topic. Therefore, it is important to provide a proper collection of tweets during a specific period of time. Unfortunately, one of the biggest problems of Twitter is the inability to provide access to the past tweets for account owner as well

[1] `https://twitter.com/whitehouse`, Accessed: 21 August 2015

as for all other users. This problem is described in more detail in chapter 2. The solution for saving tweets for further analysis was introduced in Thomas Altmann's master's thesis [Altmann, 2014], where the tweets and hashtags for a particular event have been stored and prepared for further analysis. Raw tweets are not very helpful for any kind of analysis and they need to be processed and properly visualized in order to achieve the maximum results. This thesis describes the implementation and visualization of those tweets.

1.1. Research Objectives

Since Twitter is a real-time communication social network, we need to store data permanently and analyze it. Once a particular event has passed, it is very difficult to find all the tweets corresponding to the event, and to analyze and visualize them at the same time. A common user always wants to find out the most important information about an event, such as all the posted HTTP links, users who tweet the most, most popular hashtags, images, and so on. In such cases, there is a necessity for a tool that can provide such information for the common user after the event is over. Therefore, a question can be stated out: What kind or sort of information are we capable of providing during and after Twitter event?

1.2. Structure of the Thesis

In addition to the programming terms, Chapter 2 gives an overview of terms considering social networks, micro-blogging, and twitter-based terminology such as tweet, retweet, mentions and so on.

State of the art in Chapter 3 describes everything about the scientific research on topics including Twitter and Twitter based data. It also covers some useful software solutions for data analyzing or visualization.

Chapter 4 describes approach, concept, implementation and the usage of a tool called TwitterSuitcase which was created in this master's thesis.

The evaluation of implementation is discussed in Chapter 5. This process includes interpretation of the results received in Use-Cases.

The future of Twitter, Twitter data and mostly the visualization of that data is discussed in Chapter 6.

Chapter 7 provides the final conclusion considering this master's thesis.

2. Definitions and Terminology

The aim of this chapter is to provide sufficient information for a reader considering various terms and definitions used in this thesis.

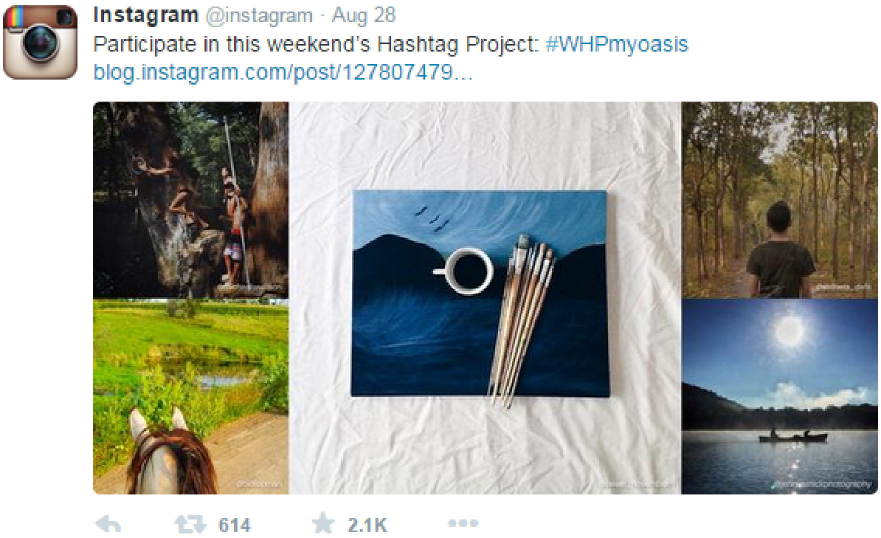
2.1. Social Network and Networking

Social network is defined as a website that gathers a group of people in order to share ideas, interests, make new friendships or just a talk. Although this term was coined before the Internet appeared, it is quite usual that nowadays it is closely connected to the Internet era. Social networking derives from the social network and represents a group of individuals gathered into a specific group (for example a group of teachers) sharing specific information and following or supporting the same objectives [Richter et al., 2011].

Typical examples of social networks are Facebook, Twitter, Youtube, LinkedIn, Pinterest and many more.

2.2. Micro-Blogging

Micro-Blogging is a light version of blogging regarding mainly the length and the type of the content. The content of the posts in the micro-blogs is typically restricted to a maximum of 200 characters (in the case of Twitter, it is 140 characters) and they mainly consists of short sentences, HTTP links or images [Kaplan and Haenlein, 2011].



Instagram @instagram · Aug 28

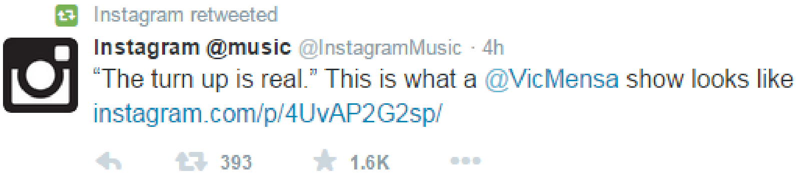
Participate in this weekend's Hashtag Project: #WHPmyoasis blog.instagram.com/post/127807479...

Figure 2.1.: Public tweet from Instagram profile

2.3. Twitter

Twitter is one of the most popular micro-blogging services in the world with approximately 316 million active users per month and over 500 million tweets sent per day[1]. Twitter's default setting for tweets is public, which gives other users an opportunity to subscribe to future tweets of an account by "following" the mentioned account. Those users are called "followers" and the subscribed user is called "followees". On the contrary, if the account settings are changed to be non-public, then the confirmation from the account owner will be needed in order to subscribe to the future tweets [O'Reilly and Milstein, 2011].

[1] https://about.twitter.com/company, Accessed: 21 August 2015



Instagram retweeted

Instagram @music @InstagramMusic · 4h

“The turn up is real.” This is what a @VicMensa show looks like instagram.com/p/4UvAP2G2sp/

Figure 2.2.: Example of a Retweet

2.3.1. Tweet

A tweet is a posted message within Twitter. It has a maximum length of 140 characters and the content can include random thoughts, HTTP links, personal messages, images or anything that fits within the characters requirements. A simple example of a tweet is shown in Figure 2.1 with text and photo posted from Instagram's[2] profile on Twitter.

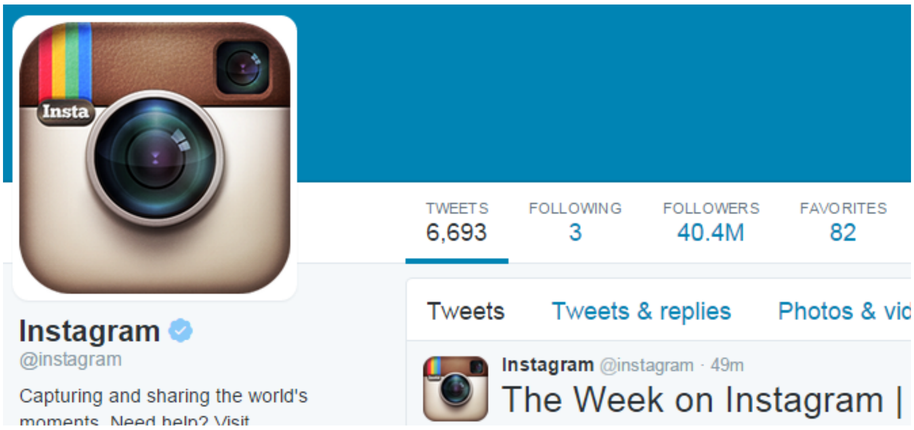
2.3.2. Retweet

A retweet is publishing a post that has already been published by someone else, to spread the information among your own Twitter followers. Basically, there are two ways to retweet by using the "retweet" button that Twitter provides or simply by adding the abbreviation for retweet "RT" including the username of the original author that you are publishing from. The Figure 2.2 shows an example of a retweeted message. Even though there is a set of rules on how to retweet properly, many problems still arise due to the different retweeting styles which makes it harder to precisely distinguish between the original tweets and retweets [Boyd et al., 2010].

2.3.3. Hashtag

Hashtag is a synonym for a "#" character, but within the context of Twitter, it represents a whole world, or more precisely a tag, which starts with "#". An

[2] https://instagram.com/, Accessed: 21 August 2015



TWEETS	FOLLOWING	FOLLOWERS	FAVORITES
6,693	3	40.4M	82

Instagram
@instagram

Capturing and sharing the world's moments. Need help? Visit

Instagram @instagram · 49m
The Week on Instagram |

Figure 2.3.: Number of tweets, following other users, followers and favorites for Instagram users

appropriate example of a hashtag would be *#bluesky*. The usage of a hashtag gives the user a possibility to associate his or her tweet with a certain topic. A single tweet can contain multiple hashtags. Twitter interprets the hashtags in tweets as tags and links them to a specific subject or topic, so they can be easily found by other users.

2.3.4. Followers

Followers are Twitter users who have subscribed to follow a specific user and see his or her tweets on their timeline. Figure 2.3 shows an example of the number of followers of an Instagram user.

2.3.5. Following

Following gives an insight into the number of users that have been subscribed to or are followed by the current user, as is shown in Figure 2.3.

2.3.6. Mention

A mention is a regular tweet, which contains another user's username within the 140 characters. The mention should be used in "@username" form. Retweets are also mentions due to the form already explained in chapter 2.3.2.

2.4. URL Shortening

The URL Shortening is a technique of shortening the uniform resource locator (URL) by reducing its length without changing the location itself. With the propagation of social networks including Twitter, where messages must fit into 140 characters, reducing the length of URLs became crucial [Neumann et al., 2010]. The maximum length of a short URL does not exceed 30 characters. There is a large number of URL shortening services (also known as USS) but the most popular ones are TinyURL[3], Goo.gl[4], Bitly[5], Bit.do[6] and many more. These services are often called URL Shorteners.

2.5. API

API is an acronym for Application Program Interface and represents a set of routines, programming instructions, standards and tools providing access to specific functions in order to build a software application. The API is responsible for the communication between software components and can be used in different parts of implementation. The most known APIs are YouTube API[7], GoogleMaps API[8], Twitter API[9] and many more.

[3] http://tinyurl.com/, Accessed: 21 August 2015
[4] https://goo.gl/, Accessed: 21 August 2015
[5] https://bitly.com/, Accessed: 21 August 2015
[6] http://bit.do/, Accessed: 21 August 2015
[7] https://developers.google.com/youtube/, Accessed: 21 August 2015
[8] https://developers.google.com/maps/, Accessed: 21 August 2015
[9] https://dev.twitter.com/, Accessed: 21 August 2015

2.6. XML

XML stands for EXtensible Markup Language and it is a markup language like HTML[10] designed not to display but to describe data. XML sets a precondition to define your own tag and is a self-descriptive. It is recommended by the World Wide Web Conssortium (W3C)[11] for consistent data exchange between individuals or companies.

2.7. JSON

JSON is an abbreviation for JavaScript Object Notation[12] and is a way of storing information as structured data. The primary usage of JSON is found within the asynchronous client-server communication, mainly replacing the XML as a data format. Although derived from JavaScript, it is language- and platform-independent data format, which makes it suitable for working with various programming languages and operating systems.

2.8. HTTP

HyperText Transfer Protocol (or short HTTP) is a stateless protocol used by World Wide Web for defining how the messages are transmitted and formatted. It also defines what sort of action a web server should execute depending on the input information. The protocol is considered to be stateless because it does not have any knowledge about previously executed commands and is executed independently.

[10] http://wiki.selfhtml.org/wiki/HTML, Accessed: 21 August 2015

[11] http://www.w3.org/, Accessed: 21 August 2015

[12] http://json.org/, Accessed: 21 August 2015

2.9. PHP

Originally, PHP stood for ”Personal Home Page”, but over the years the meaning of the abbreviation has changed to ”Hypertext Preprocessor” which is considered to be a recursive backronym. It is a very popular scripting language designed specifically for web development and can easily be combined with ”static” HTML.

2.10. RSS

Rich Site Summary (RSS) is a standard web format used to frequently publish information including text and metadata (date and time of publishing and author’s name). The published information could include headlines, blog entries or media data (such as video and audio data).

2.11. GPS

Global Positioning System (GPS) is a navigation system based on a group of satellites placed into the Earth orbit which can provide the exact location and time information all over the world at any time of a day.

3. State of the Art

This chapter gives an overview of the scientific researches and a short insight into the available tools for research and analysis of Twitter data.

3.1. Scientific Research

Only one year after Twitter was launched, a group of researchers have published one of the first analyses of Twitter data and have published it under the title "Why We Twitter: Understanding Microblogging Usage and Communities" . The first objective of this paper was to study topological and geographical properties of Twitter's network. The second objective was to analyze the user intentions in combination with community level and finally to show how the users with the same intentions interact with each other [Java et al., 2007]. The results have shown that Twitter user can be divided into four major groups, in accordance with their intentions:

- *Daily chatter* : The largest and most common group of users of Twitter. The majority of their post is based on a daily routine.
- *Conversations* : Replies to the posts of other users since there is no direct way to comment. Approximately 21% of users belongs to this group.
- *Sharing information/URLs* : About 13% of all posts contains some URL.
- *Reporting news* : A group of users who are frequently posting news or comments about current events on Twitter. This is very popular among different services, such as weather forecast, since Twitter has enabled an easy access to the developer API.

The user himself can be distinguished into three main categories:

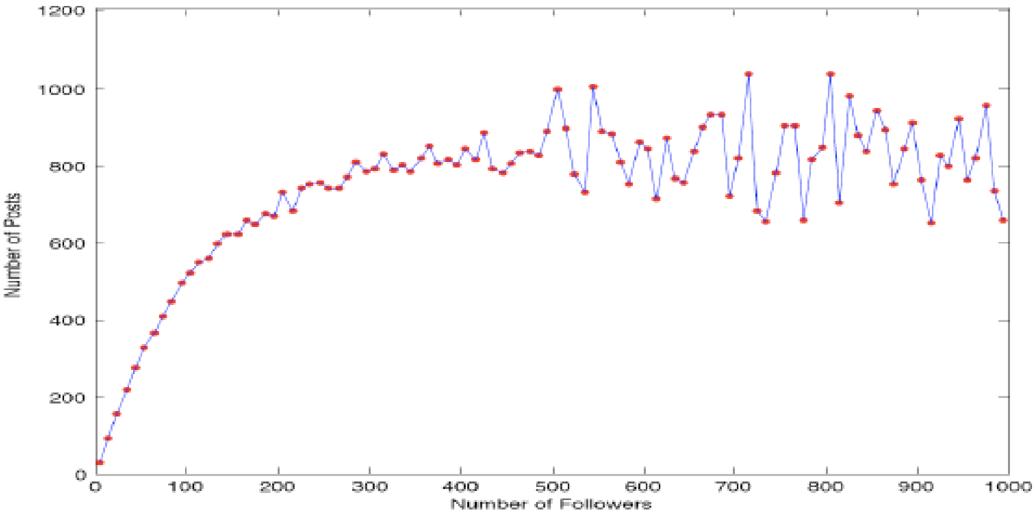


Figure 3.1.: Number of posts as a function of the number of followers, [Huberman et al., 2008], P. 4.

- *Information Source* : This group of users is well known for its valuable updates, which makes them very popular. The updates may be posted regularly or infrequently. This group has gained popularity due to the quality and valuable nature of the content.
- *Friends* : The widespread group of users with additional sub-categories, such as family or co-workers.
- *Information Seeker* : An information seeker posts rarely, but follows other users regularly.

The research has also revealed a users with multiple intentions. This type of users could make other users feel overwhelmed by Twitter, since there is no categorization of tweet nor that would distinguish between information that should and information that should not be seen. This resulted in a recommendation categorizing friends into specific groups such as family, co-workers or closest friends [Lavallee, 2007].

In the research published under the title ”Social Network That Matters”, the researchers distinguished a hidden network of connections which underlies the ”declared” set of friends and followers [Huberman et al., 2008]. The

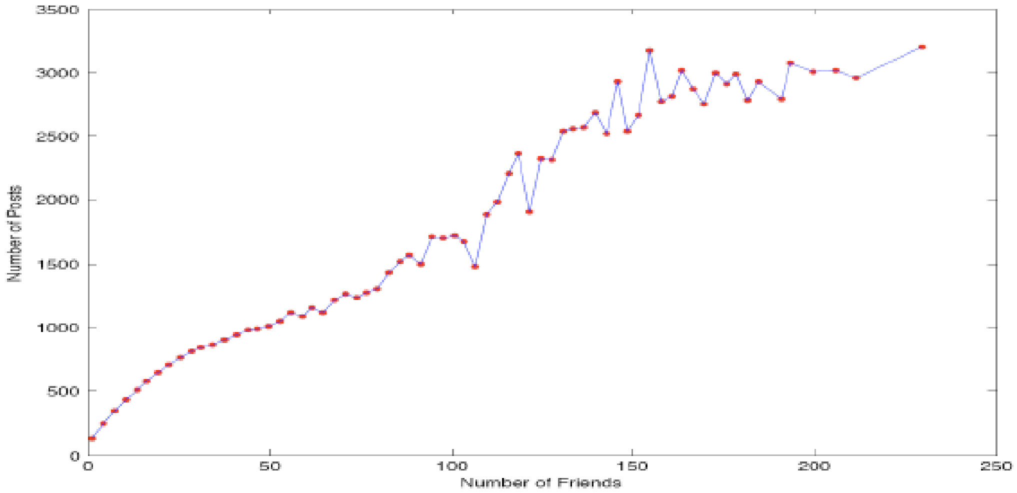


Figure 3.2.: Number of posts as a function of the number of followers, [Huberman et al., 2008], P. 4.

term a user's *friend* is basically defined as a person who the user has directed as least two posts to (using mentions). The number of posts initially increases as the number of followers increases but it eventually saturates, as shown in Figure 3.1. As the number of friends grows, the number of posts increases as well, reaching 3200 without saturation. The limit of 3200 represents the Twitter's limitation of displaying updates per user. The relation between the number of posts and number of friends is shown in Figure 3.2.

The results have shown that the majority of Twitter users have a small group of friends with whom they interact regularly. There is also a larger group of users that they have subscribed to due to direct interest in the content of their status updates.

Using a large amount of collected data from Twitter, the researchers in [Cha et al., 2010] have tried to measure the user's dynamic influence across time and topics within Twitter. The comparison was based on three different categories or measures of influence like indegree (refering to number of followers), retweets and mentions. The underlying reasons for the choice of measurements were that indegree determines the popularity of a user, retweets indicates the ability of the user to create valuable content and mentions determines the ability of that user to engage others in a conversation. The overlap of the results is shown in Figure 3.3. The results have shown that the most followed users are various public figures and news sources. Such users are Barack Obama (politician), Shaquille O'Neal (basketball player), Britney Spears (singer), New York Times (newspaper), CNN (television channel), and so on. The conclusion of this research was that popularity does not automatically lead to an influence. Although it does provide a better position, due to the number of followers, a user needs to provide a great personal effort in order to gain influence.

[Jansen et al., 2009] have conveyed a research on Twitter, a microblogging service, as a form of electronic word-of-mouth in correlation to brands and the influence of the service on various brands. The Summize[1] service was used to analyze the collected data which resulted in classification of the brands into five different groups from lowest to highest as:

- *No Sentiment* : Tweet without emotions or special punctation, containing just brand mention.
- *Wretched* : Tweet gives a generally negative connotation to a brand.
- *Bad* : Tweet consists mostly of negative phrases and words, but there could be a few positive opinions or statements.
- *So-so* : The number of positive and negative statements is almost equal. It is difficult to determine whether the statement is positive or negative.
- *Swell* : Tweet is mainly positive, but there could be some light-weighted negative phrases.
- *Great* : Clearly the most positive sentiment. Here, the account of negative words, phrases and statements is the lowest.

The test period of over 131 weeks for 50 brands gave approximately 149,472 tweets and the results can be seen in Figure 3.4.

The brands have been structured according to the industry sector such as:

- *Apparel* : Banana Republic, H&M, TopShop

[1] Summize was aquired by Twitter in August of 2008 and is no longer available as as an independent service

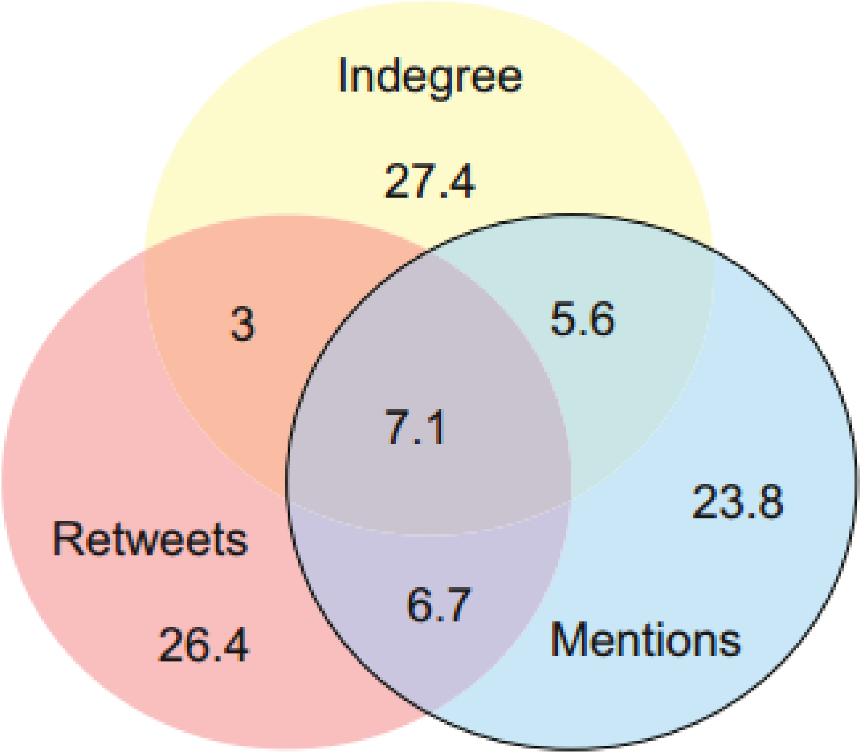
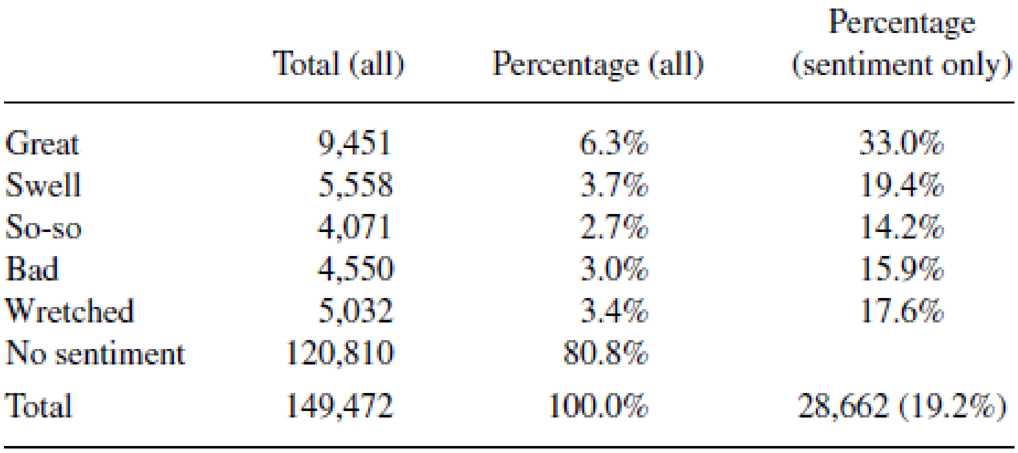


Figure 3.3.: Venn normalized diagram of the top 100 influentials across measures, [Cha et al., 2010], P. 4.



	Total (all)	Percentage (all)	Percentage (sentiment only)
Great	9,451	6.3%	33.0%
Swell	5,558	3.7%	19.4%
So-so	4,071	2.7%	14.2%
Bad	4,550	3.0%	15.9%
Wretched	5,032	3.4%	17.6%
No sentiment	120,810	80.8%	
Total	149,472	100.0%	28,662 (19.2%)

Figure 3.4.: Analysis of individual tweets for sentiment., [Jansen et al., 2009], Table 4, P. 10.

- *Automotive* : Honda, Mini Clubman, Prius, Smart ForTwo, Toyota
- *Computer Hardware* : Dell, Lenovo, MacBook Air
- *Computer Software* : Microsoft, Leopard
- *Energy* : Exxon, Sunoco
- *Fast Food* : McDonald's, Starbucks
- *Food* : Kellogg's, Malt-O-Meal
- *Internet Service* : Amazon, Facebook, Gmail, Google
- *Personal Care* : Aquafresh, Oral-B
- *Sporting Goods* : Nike, Adidas, Reebok
- *Transportation* : DHL, FedEx, Forever Stamp

As one can see from Figure 3.4, the majority of the mentions in tweets about some brand expressed no sentiment. Nevertheless, the remaining 19,2% should not be underestimated and can definitely have influence on brand or products. This basically means that common users are using Twitter for general information and information-seeking, for asking questions and sharing information about various brands or products.

As stated by [Zhao and Rosson, 2009], microblogging provides a new communication channel for people to publicly broadcast personal information or information in general which they would not publish using existing chan-

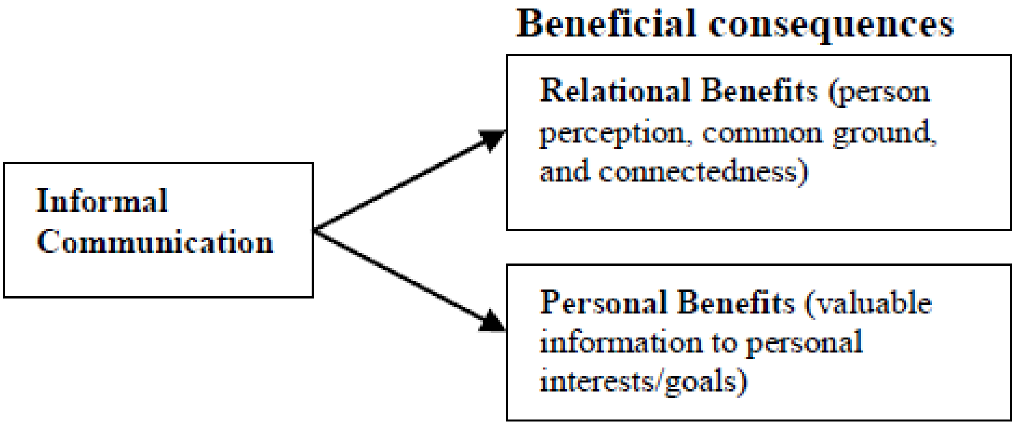

Beneficial consequences
Informal Communication
Relational Benefits (person perception, common ground, and connectedness)
Personal Benefits (valuable information to personal interests/goals)

Figure 3.5.: Proposed benefits of informal communication., [Zhao and Rosson, 2009], Figure 1, P. 2.

nels, such as phone, email, weblog, and so on. This research considers the usage and potential impacts of microblogging on informal communication at work. They have organized the benefits of informal communication into relational and personal beneficial consequences (Figure 3.5).

The relational consequences consist of a person's perception (of other persons), developing common ground and feeling of connectedness. The personal consequences are wrapped around the personal interests and goals. The data in the study was obtained through a phone interview with participants from a large IT company. The range of the participants included workers, managers in engineering, marketing, product management and corporate communication. During the analysis of the interviews, the researchers have noticed some differences in the way the interviewees see and feel about Twitter. Those differences have been divided into three opinions:

- Frequent brief updates about personal life activities
- Real-time information
- People-based RSS feed

The technology features have also been structured and divided into three characteristics:

- Brevity
- Mobility and pervasive access
- Broadcast nature

The results have shown that the ”Work-relevant information sharing and expertise seeking” are highly appreciated and are leading to positive effects within the personal impacts. One of the interviewees stated that the possibility of following unknown people from different companies who work in the same or equal department gives them a broader perspective. However, there is also a certain amount of risk when publishing company-related information which is publicly accessible. Furthermore, there are some separation issues between work-related and private updates on Twitter.

Microblogging services are valuable sources of data when it comes to opinion mining and sentiment analysis. The research done by Pak and Paroubek in [Alexander Pak, 2010] focuses on sentiment analysis within the Twitter. The objective of the research was automatic collection of data for sentiment analysis and opinion mining purposes. The crucial method was classifying the collected tweets into three main categories such as positive (containing happy emoticons), negative (containing sad emoticons) and neutral (no emoticons at all). The foundation stone was built around the neutral sentiment (posts from Twitter accounts such as New York Times or Washington Post), which have been used as a training data set for sentiment classifier. Since the sentiment classifier was based on the multinomial Naive-Bayes Classifier[2] using N-grams and Part-Of-Speech Tags[3] as features, they have managed to achieve very high accuracy.

The term ”Place-Triggered Geotagged Tweet” was established by Hiruta er al. in their research [Hiruta et al., 2010], where tweets contain both geotag and content-related relation to user’s location. The basic assumption of this research is that the real world will be structured as a collection of descriptive attributes. In the following step, there is a necessity for a wrapper system which can *extract*, *classify* and *provide real-time dynamic attributes* for a searched event. The focus of the research was placed on Twitter because of its public and agile nature as a communication medium. The method used

[2] http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html, Accessed: 22 August 2015

[3] http://nlp.stanford.edu/software/tagger.shtml, Accessed: 21 August 2015

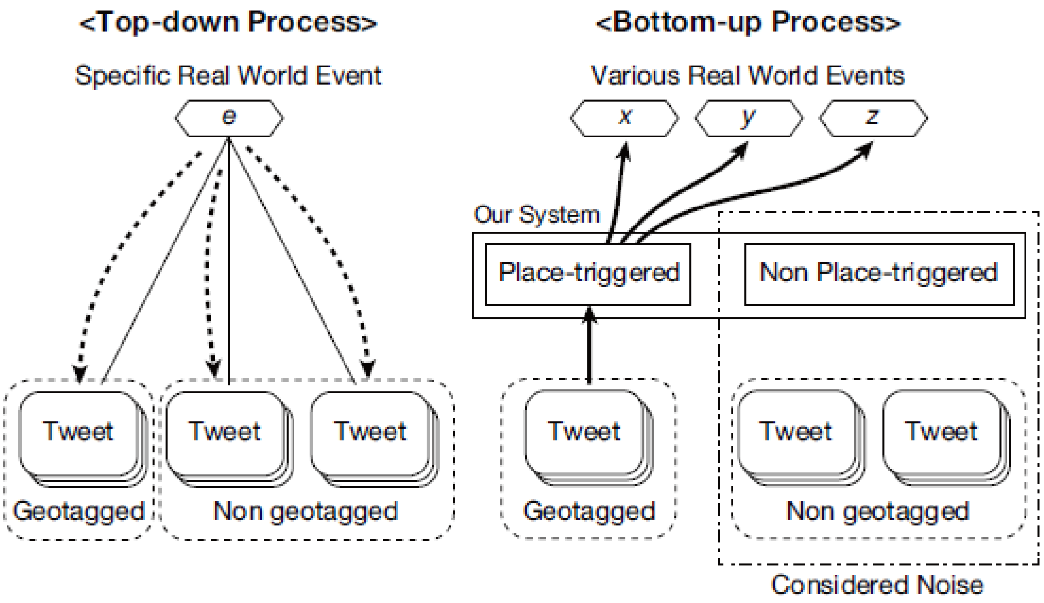


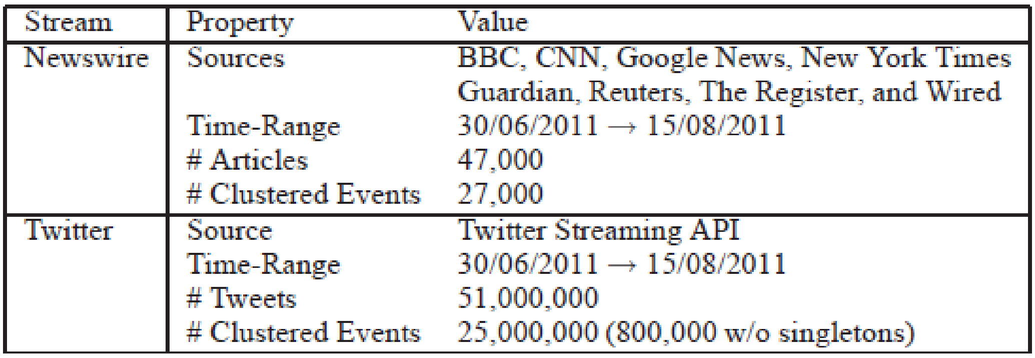
Figure 3.6.: Comparison of approaches to detect events from tweets., [Hiruta et al., 2010], Figure 1, P. 2.

can be described as Top-Down Process and Bottom-Up Process, as is shown in Figure 3.6. The objectives of this research are divided into two phases:

- *Detect* : Detection of place-triggered geotagged tweets which also determines if the tweet contains relation to the location.
- *Classify* : Classification of the place-triggered geotagged tweets by filtering the content based on keywords and regular expressions.

The tweets are classified into five types:

- *Report of whereabouts* : Tweet referring to the user's current location.
- *Food:* : Tweet sharing the information about food or drink.
- *Weather:* : Tweet about the weather at the location.
- *Back at home* : Tweet about straightforward information - that the entity (mostly a person) came back home.
- *Earthquake:* : Tweet containing information about an earthquake.



Stream	Property	Value
Newswire	Sources	BBC, CNN, Google News, New York Times Guardian, Reuters, The Register, and Wired
	Time-Range	30/06/2011 → 15/08/2011
	# Articles	47,000
	# Clustered Events	27,000
Twitter	Source	Twitter Streaming API
	Time-Range	30/06/2011 → 15/08/2011
	# Tweets	51,000,000
	# Clustered Events	25,000,000 (800,000 w/o singletons)

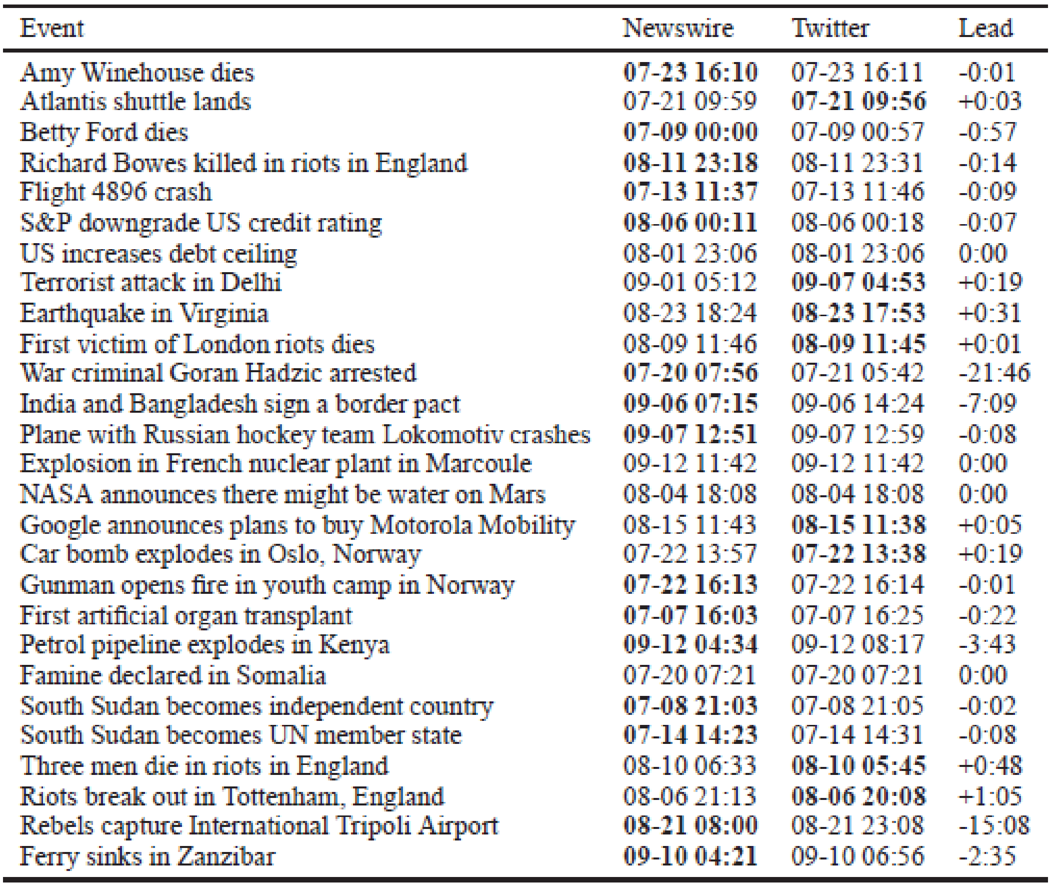
Figure 3.7.: Stream Statistics., [Petrovic et al., 2013], Table 1, P. 2.

Though this research is experimental and has not been implemented yet, the results obtained from 18 human classifiers are quite encouraging as they have achieved accuracy of 82%.

Twitter is very often considered to be a powerful source of real-time data, but does it have a potential to replace traditional news-wires? Petrovic et al. in the research [Petrovic et al., 2013] have tried to analyze this question and gives an overview of the common aspects and overlapping areas between Twitter and news-wires. The analysis was performed by manual identification of news both on Twitter and news events. The duration of the process was set to a period of 2 months. The reports of the statistics are shown in Figure 3.7. An overview of the performance test was also created in Figure 3.8, including Event title, news-wire time of publishing, Twitter time of publishing and the difference between the two (measured in seconds).

The results of the study have shown that almost all news provided by a news-wire were covered on Twitter and that a number of events reported on Twitter were not mentioned in news-wire. Nevertheless, it is not possible to say whether Twitter or News-wires are more accurate or faster.

Working with microblogging services has turned out to be more interesting, like [Ebner and Maurer, 2008] have noticed in the lecture ”Social Aspects of Information Technology”. In order to increase a student’s activity they had to consider three crucial didactical factors:



Event	Newswire	Twitter	Lead
Amy Winehouse dies	**07-23 16:10**	07-23 16:11	-0:01
Atlantis shuttle lands	07-21 09:59	**07-21 09:56**	+0:03
Betty Ford dies	**07-09 00:00**	07-09 00:57	-0:57
Richard Bowes killed in riots in England	**08-11 23:18**	08-11 23:31	-0:14
Flight 4896 crash	**07-13 11:37**	07-13 11:46	-0:09
S&P downgrade US credit rating	**08-06 00:11**	08-06 00:18	-0:07
US increases debt ceiling	08-01 23:06	08-01 23:06	0:00
Terrorist attack in Delhi	09-01 05:12	**09-07 04:53**	+0:19
Earthquake in Virginia	08-23 18:24	**08-23 17:53**	+0:31
First victim of London riots dies	08-09 11:46	**08-09 11:45**	+0:01
War criminal Goran Hadzic arrested	**07-20 07:56**	07-21 05:42	-21:46
India and Bangladesh sign a border pact	**09-06 07:15**	09-06 14:24	-7:09
Plane with Russian hockey team Lokomotiv crashes	**09-07 12:51**	09-07 12:59	-0:08
Explosion in French nuclear plant in Marcoule	09-12 11:42	09-12 11:42	0:00
NASA announces there might be water on Mars	08-04 18:08	08-04 18:08	0:00
Google announces plans to buy Motorola Mobility	08-15 11:43	**08-15 11:38**	+0:05
Car bomb explodes in Oslo, Norway	07-22 13:57	**07-22 13:38**	+0:19
Gunman opens fire in youth camp in Norway	**07-22 16:13**	07-22 16:14	-0:01
First artificial organ transplant	**07-07 16:03**	07-07 16:25	-0:22
Petrol pipeline explodes in Kenya	**09-12 04:34**	09-12 08:17	-3:43
Famine declared in Somalia	07-20 07:21	07-20 07:21	0:00
South Sudan becomes independent country	**07-08 21:03**	07-08 21:05	-0:02
South Sudan becomes UN member state	**07-14 14:23**	07-14 14:31	-0:08
Three men die in riots in England	08-10 06:33	**08-10 05:45**	+0:48
Riots break out in Tottenham, England	08-06 21:13	**08-06 20:08**	+1:05
Rebels capture International Tripoli Airport	**08-21 08:00**	08-21 23:08	-15:08
Ferry sinks in Zanzibar	**09-10 04:21**	09-10 06:56	-2:35

Figure 3.8.: Times (in UTC) of events, first newswire stories, first tweets and lead (+ when Twitter leads)., [Petrovic et al., 2013], Table 4, P. 3.

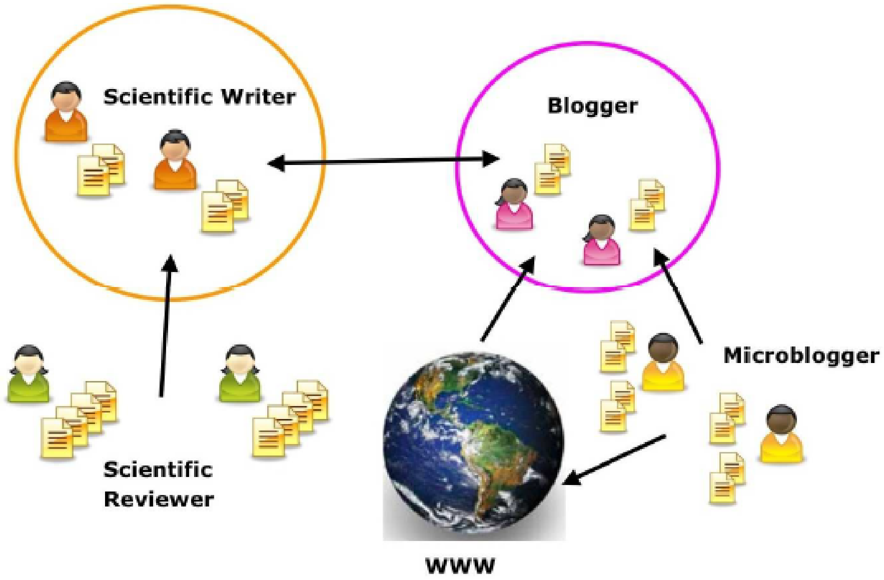


Figure 3.9.: Overview of the didactical concept., [Ebner and Maurer, 2008], Figure 1, P. 5.

- *Reflection* : Evaluation of the experts' presentation in order to form own opinion.
- *Discussion* : Enhance the visible spectrum.
- *Quality* : Ensure that arguments and opinions are based on scientific methods, rules and approaches.

During the course, students have been divided into one of the following four groups with different tasks:

- *Scientific Writer* : Student has to write two short articles on a topic of their own choice.
- *Scientific Reviewer* : Review of the articles from the first group.
- *Blogger* : Maintenance of the lecture's blog by writing at least two weblog-posts each week.
- *Microblogger* : This group needs to post at least two microblogs each week and also needs to comment on at least two blog posts, created by a third group on a topic of their own choice.

The overview of the concept is presented in Figure 3.9.

To support the theory that Twitter could be applied into any sphere of life, the researchers in [Sakaki et al., 2010] have tested the possibility to use real-time data in order to notify the public about the earthquake. They have designed a system which classifies the given data in order to determine the exact location of places struck by earthquakes at that moment and gives a global warning in the form of notifications for impacted areas. Since Japan has numerous earthquakes and a large number of Twitter users who are spread across the entire country, it has been selected as an optimal test region. The assumption of the researcher was that every Twitter user is regarded as a *sensor* and each tweet is considered to be *sensor information* also called *social sensors*. Due to the classification of the tweets, the researchers distinguish between relevant and irrelevant tweets. Relevant tweets are considered to be tweets containing location data, either in the form of the GPS or in the content of the tweet. The experimental model of this research proved to be very useful during the test phase in 2009 and was finally developed as an Earthquake Reporting System called ”Toretter[4]” which has been fully operational since August, 2010. The latest results are indicating that the ”Toretter” notifies users and alerts them faster then the Japan Meteorological Agency.

The use of microblogging services was discussed in the research of Ebner et al. and published in ”How People are using Twitter during Conferences” [Ebner et al., 2008]. They have analyzed how the use of a special hashtag for Twitter before, during and after the conference can be exploited, what motives lie behind tweeting during a conference and finally what value that information carries. The use of Twitter was divided into three different stages of a conference such as Before, During and After a conference. The participants of the study had to complete a survey containing 34 questions. The subjects were required to answer if they already had a Twitter account, if they were using it for professional or private purposes or both, if they are using Twitter to actively communicate during conferences, and so on. The results have shown that 95.1% of users already had a Twitter account and they have been using it for both personal and private reasons. What was interesting about the results is that about 51% of users was applying the same approach of ”single account for multiple use cases” in other communication tools. Over two-thirds or 67.5% of Twitter users tweeted

[4] It means ”we have taken it” in Japanese.

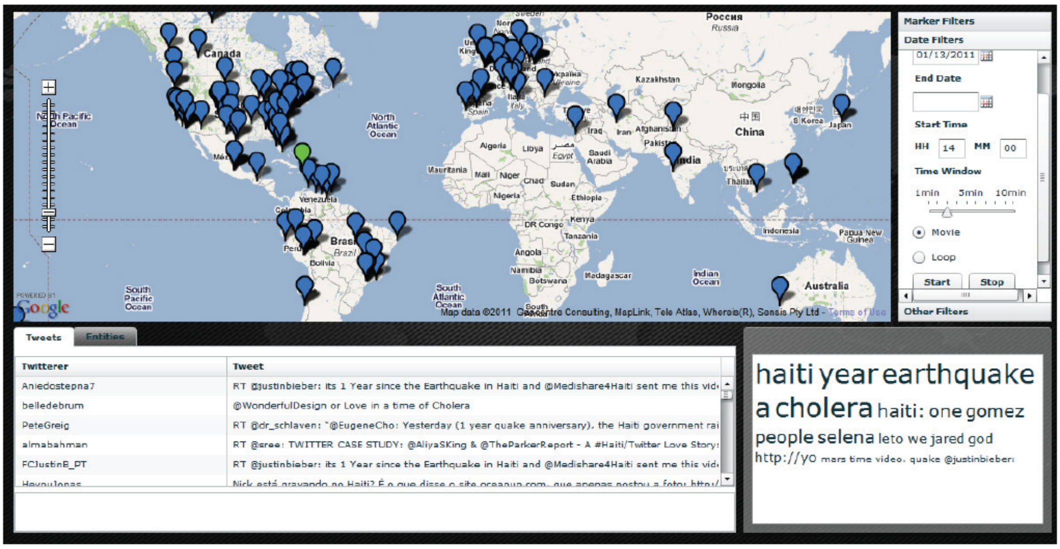
actively during a conference. The content of tweeted text was also analyzed and the majority of the content or 50% of it was defined as a plain text (without any links or images). Approximately 10% of the tweets contained links to external services and were mainly sent by the conference delegates. The survey also included questions about expectations and attitudes towards using Twitter during a conference. The users were quite sceptical at first, but their attitude towards the use of Twitter changed through discussion, spreading and sharing conference-related information. The final results of this research have shown that discussion on various topics is not limited only to the face-to-face audience and could be exploited effectively. There were also some disadvantages like impracticality to work with data formats other than plain text and web links.

3.2. Available Tools

This sections gives a brief overview of the existing tools which can be used for data analysis. Since the number of tools can not be precisely defined, the overview of tools will cover only the most popular ones.

3.2.1. TweetTracker

TweetTracker was explicitly designed for monitoring and analyzing relevant tweets from different perspectives. An elaborate explanation of the TweetTracker in a case study of Cholera outbreak in Haiti was explained in [Kumar et al., 2011]. It collects data according to content, hashtags, URLs, mentions, time and location. A variety of visualization for collected data is allowed which include the projection of tweets on various geographical maps, automatic translation of Non-English tweets and keyword trending with comparison. The visualization is allowed for the 7500 tweets which are first collected and it supports export of a data into formats like XML or JSON as described in chapters 2.6 and 2.7. The Figure 3.10 displays the visualization of tagclouds. TweetTracker does not fetch all the data due to the produced amount of Twitter's data, which is approximately 20MB per second. Although not all data is fetched, the results are satisfying because it



01/13/2011

End Date

Start Time: HH 14 MM 00

Time Window: 1min 5min 10min

● Movie

○ Loop

Start Stop

Twitterer	Tweet
Aniedostepna7	RT @justinbieber: its 1 Year since the Earthquake in Haiti and @Medishare4Haiti sent me this vid
belledebrum	@WonderfulDesign or Love in a time of Cholera
PeteGreig	RT @dr_schlaven: "@EugeneCho: Yesterday (1 year quake anniversary), the Haiti government rai
almabahman	RT @sree: TWITTER CASE STUDY: @AliyaSKing & @TheParkerReport - A #Haiti/Twitter Love Story:
FCJustinB_PT	RT @justinbieber: its 1 Year since the Earthquake in Haiti and @Medishare4Haiti sent me this vid
HeyouJonas	Nick está gravando no Haiti? É o que disse o site oceanup.com, que apenas postou a foto: http:/

haiti year earthquake a cholera haiti: one gomez people selena leto we jared god http://yo mars time video. quake @justinbieber:

Figure 3.10.: Tagclouds analysis., [Kumar et al., 2011], Figure 3, P. 2.

usually produces between 10.000 and 50.000 tweets per hour, depending on the data stream flow.

3.2.2. Tweet Sentiment Visualization

Tweet Sentiment Visualization[5] was created at the North Carolina State University[6] and it uses various ways of tweet visualization. Collections of tweets can be grouped by topic, by sentiment and by frequent terms. Single tweets are represented as a circle of different color, size, brightness and transparency which actually responds to the significance of the tweet. In Figure 3.11 we can see an example of sentiment visualization for a keyword *#football*. Color of the tweets determines if the tweet was pleasant (green) or unpleasant (blue). The more active the tweets are, the brighter they will be. Large and less transparent circles represent more confident evaluation. There

[5] http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/, Accessed: 23 August 2015

[6] http://www.csc.ncsu.edu/, Accessed: 23 August 2015

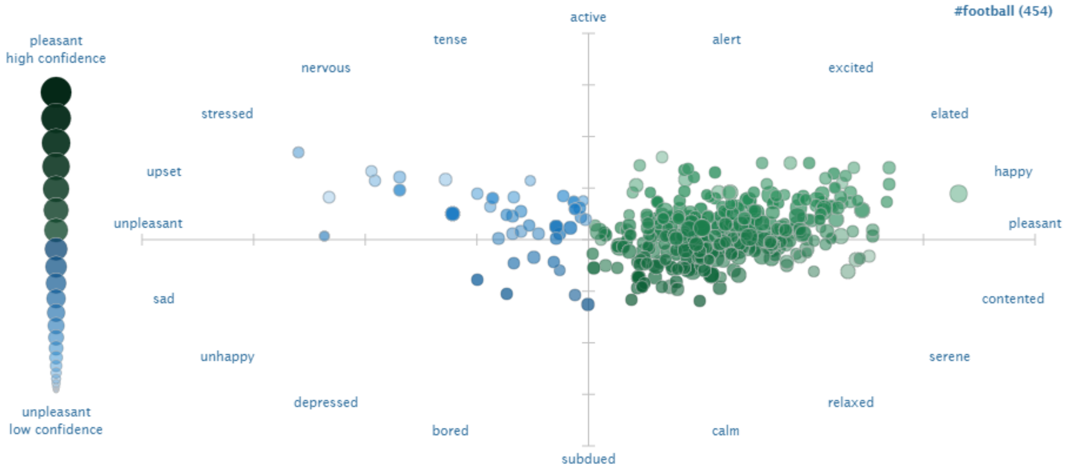


Figure 3.11.: Sentiment Visualization, Source `http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/`, Accessed: 25 August 2015.

are also other visualization methods such as clustering (see Figure 3.12) or displaying the result across the timeline of tweets (see Figure 3.13).

3.2.3. Tweet Archivist

Tweet Archivist[7] is a paid service for archiving and analysis of tweets. As is shown in Figure 3.14 the service can analyze and provide information about:

- Top Users
- Top Words
- Top URLs
- Source of Tweet
- Language
- Volume Over Time
- User Mentions
- Hashtags

[7] `http://www.tweetarchivist.com/`, Accessed: 23 August 2015

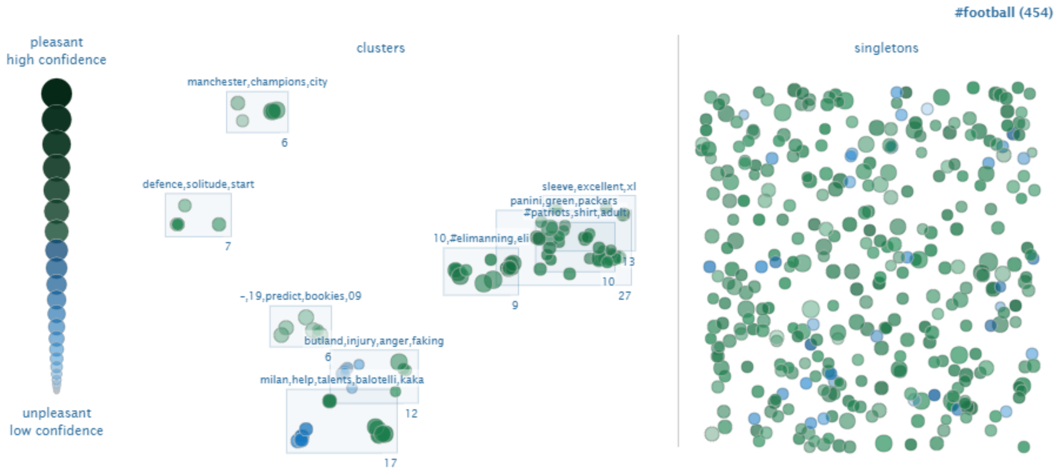


Figure 3.12.: Clustering Visualization, Source http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/, Accessed: 25 August 2015.

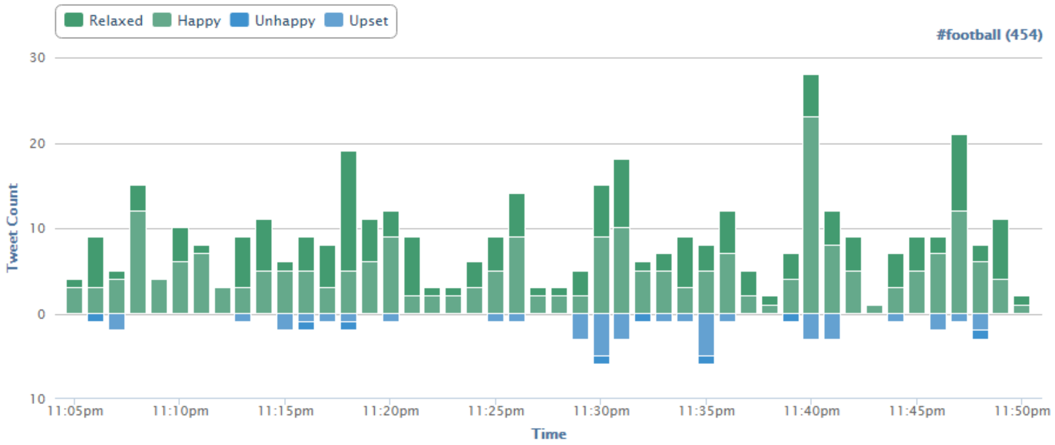
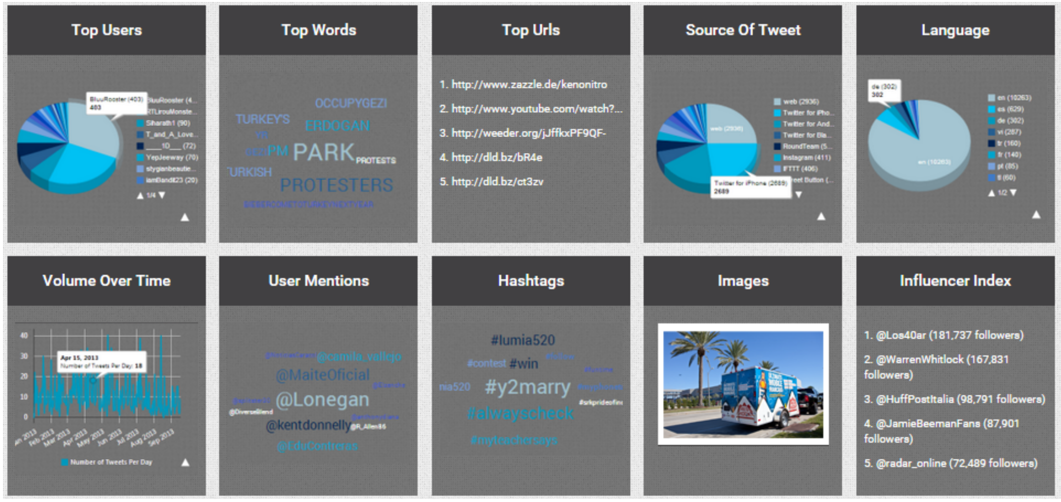


Figure 3.13.: Timeline Visualization, Source http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/, Accessed: 25 August 2015.



1. http://www.zazzle.de/kenonitro
2. http://www.youtube.com/watch?...
3. http://weeder.org/jJffkxPF9QF-
4. http://dld.bz/bR4e
5. http://dld.bz/ct3zv

1. @Los40ar (181,737 followers)
2. @WarrenWhitlock (167,831 followers)
3. @HuffPostItalia (98,791 followers)
4. @JamieBeemanFans (87,901 followers)
5. @radar_online (72,489 followers)

Figure 3.14.: Tweet Archivist, Source http://www.tweetarchivist.com/, Accessed: 23 August 2015.

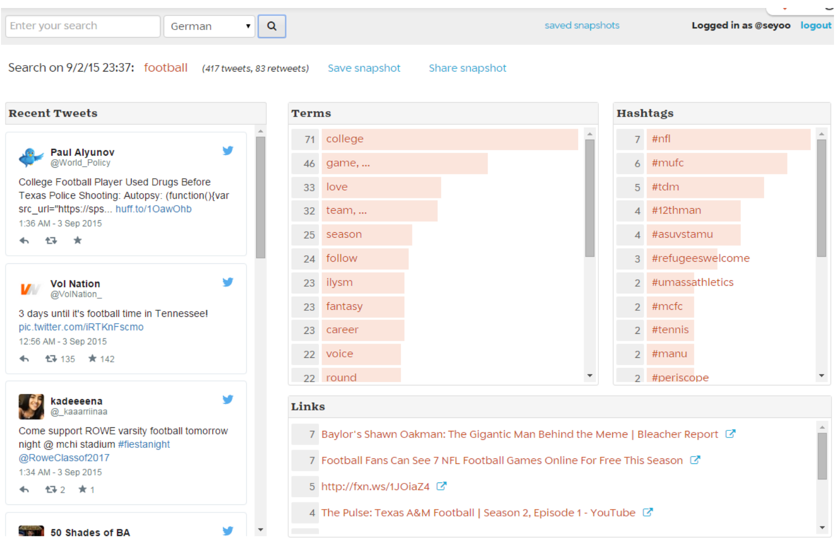
- Images
- Influencer Index

One of the disadvantages of this service is a lack of user archives and a general lack of API's which enable embedding the results elsewhere or extending the functionality for own necessities. The visualization of images is not very common among the tools dealing with Twitter data and should be considered as an advancement towards other services.

3.2.4. SocialBro

SocialBro is a paid service and his primary targets are business groups and companies. The range of SocialBro[8] products is divided into three groups: Listening & Insights, Social Marketing and Twitter Advertising. They offer a possibility to analyze data by exploring Twitter community, specify and approach the target group with appropriate advertising by matching the objectives and audience and determining the best time to tweet.

[8] http://www.socialbro.com/, Accessed: 23 August 2015



Search on 9/2/15 23:37: football (417 tweets, 83 retweets)

Recent Tweets

Paul Alyunov @World_Policy
College Football Player Used Drugs Before Texas Police Shooting: Autopsy: (function(){var src_url="https://sps... huff.to/1OawOhb
1:36 AM - 3 Sep 2015

Vol Nation @VolNation_
3 days until it's football time in Tennessee! pic.twitter.com/iRTKnFscmo
12:56 AM - 3 Sep 2015
135 142

kadeeeena @_kaaarriinaa
Come support ROWE varsity football tomorrow night @ mchi stadium #fiestanight @RoweClassof2017
1:34 AM - 3 Sep 2015
2 1

50 Shades of BA

Terms

71	college
46	game, ...
33	love
32	team, ...
25	season
24	follow
23	ilysm
23	fantasy
23	career
22	voice
22	round

Hashtags

7	#nfl
6	#mufc
5	#tdm
4	#12thman
4	#asuvstamu
3	#refugeeswelcome
2	#umassathletics
2	#mcfc
2	#tennis
2	#manu
2	#periscope

Links

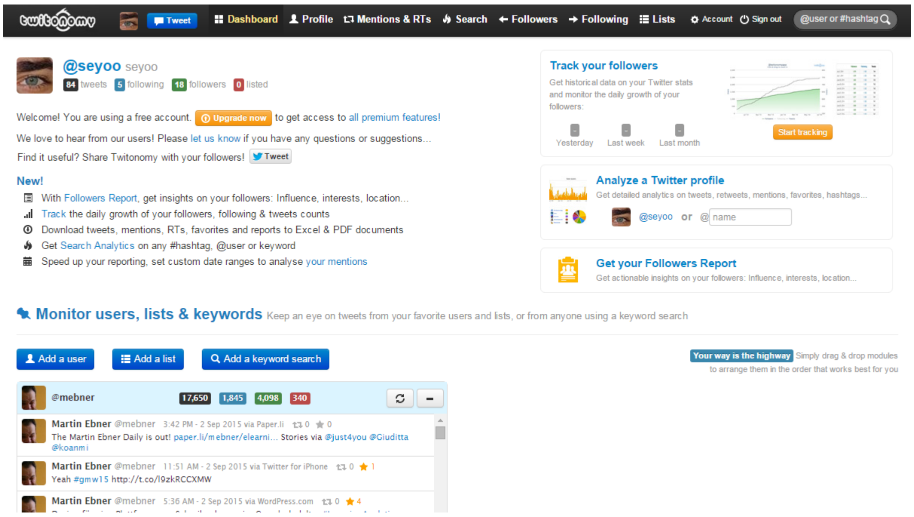
7	Baylor's Shawn Oakman: The Gigantic Man Behind the Meme | Bleacher Report
7	Football Fans Can See 7 NFL Football Games Online For Free This Season
5	http://fxn.ws/1JOiaZ4
4	The Pulse: Texas A&M Football | Season 2, Episode 1 - YouTube

Figure 3.15.: twXplorer Start page, Source https://twxplorer.knightlab.com/search/, Accessed: 24 August 2015.

3.2.5. twXplorer

twXplorer[9] was developed by the Northwestern University Knight Lab[10]. This service provides real-time analysis of tweets considering: the most used words, hashtags and top links. It also provides a search for bigrams (two-word phrases) appearing more than once, such as *white house*. There is no possibility of archiving tweets, although one can save a snapshot of the analysis for later interpretation, viewing or evaluation. The biggest handicap of this service is working only with the latest 500 tweets which makes the analysis rather limited. The overview of this tool is presented in Figure 3.15. The only requirements for using this tool is to have a Twitter account.

[9] https://twxplorer.knightlab.com/, Accessed: 24 August 2015
[10] http://www.northwestern.edu/, Accessed: 24 August 2015



@seyoo seyoo
84 tweets 5 following 18 followers 0 listed

Welcome! You are using a free account. Upgrade now to get access to all premium features!

We love to hear from our users! Please let us know if you have any questions or suggestions...

Find it useful? Share Twitonomy with your followers! Tweet

New!

- With Followers Report, get insights on your followers: Influence, interests, location...
- Track the daily growth of your followers, following & tweets counts
- Download tweets, mentions, RTs, favorites and reports to Excel & PDF documents
- Get Search Analytics on any #hashtag, @user or keyword
- Speed up your reporting, set custom date ranges to analyse your mentions

Track your followers
Get historical data on your Twitter stats and monitor the daily growth of your followers:
- Yesterday - Last week - Last month

Analyze a Twitter profile
Get detailed analytics on tweets, retweets, mentions, favorites, hashtags...
@seyoo or @ name

Get your Followers Report
Get actionable insights on your followers: Influence, interests, location...

Monitor users, lists & keywords Keep an eye on tweets from your favorite users and lists, or from anyone using a keyword search

Your way is the highway Simply drag & drop modules to arrange them in the order that works best for you

@mebner 17,650 1,845 4,098 340

Martin Ebner @mebner 3:42 PM - 2 Sep 2015 via Paper.li 0 ★ 0
The Martin Ebner Daily is out! paper.li/mebner/elearni... Stories via @just4you @Giuditta @koanmi

Martin Ebner @mebner 11:51 AM - 2 Sep 2015 via Twitter for iPhone 0 ★ 1
Yeah #gmw15 http://t.co/l9zkRCCXMW

Martin Ebner @mebner 5:36 AM - 2 Sep 2015 via WordPress.com 0 ★ 4

Figure 3.16.: Twitonomy Dashboard, Source http://www.twitonomy.com/dashboard.php, Accessed: 27 August 2015.

3.2.6. Twitonomy

This service offers two account options, free and premium. The assortment of the analysis tool in the free version is quite satisfying. Twitonomy[11] requires only a Twitter account and an approval that it can be started using Twitter authentication. Although the premium version of the service offers more tools and visualization options, the following tools or processes are available in the free version:

- user monitoring
- keyword monitoring
- analysis of the twitter profile (own and external
- most favorited tweets
- most retweeted tweets
- mentions & RTs
- followers and their analysis
- following

[11] http://www.twitonomy.com/, Accessed: 24 August 2015

- administration of your own created data

The overview of the main dashboard of Twitonomy can be seen in Figure 3.16.

3.2.7. MentionMap

Even though, MentionMap[12] was considered to be an analysis tool, it is rather a visualization tool without any analysis considering Twitter data. This tool gives a great interactive map overview of Twitter connections to other users or topics, like Figure 3.17 shows.

3.2.8. Twitter Counter

Twitter Counter[13] is also one of the existing services analyzing Twitter data and representing the results with visualization effects. The presence of a Twitter account is mandatory in order to analyze data properly. The structure of the function is more similar to Twitonomy 3.2.6 and here we distinguish between free and premium version if the full capacity of the tool is considered. Figure 3.18 gives an overview of this tool.

[12] http://mentionmapp.com/, Accessed: 25 August 2015

[13] http://twittercounter.com/, Accessed: 25 August 2015

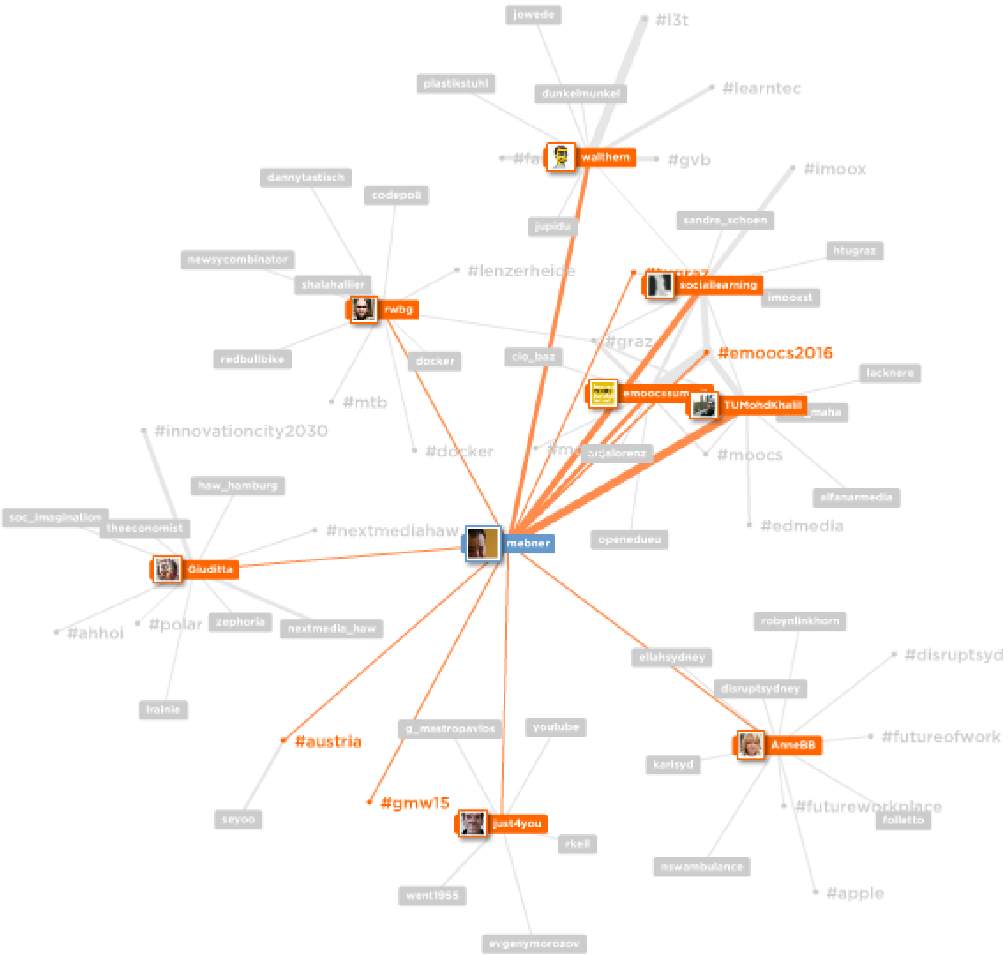
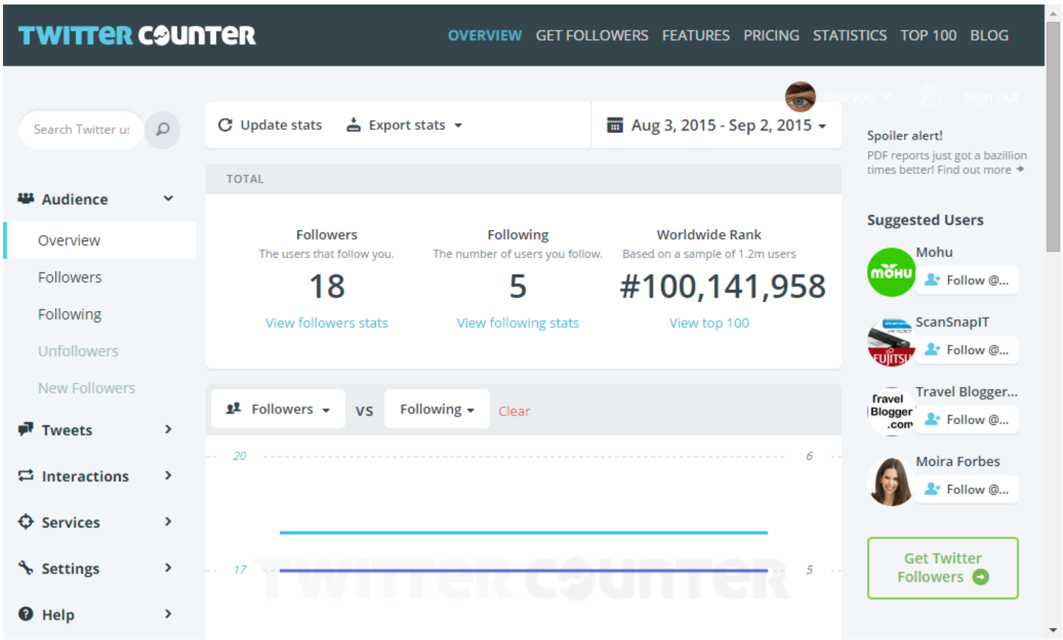


Figure 3.17.: MentionMap of the user *mebner*, Source `http://mentionmapp.com/`, Accessed: 25 August 2015.



Spoiler alert!

PDF reports just got a bazillion times better! Find out more

Followers	Following	Worldwide Rank
The users that follow you.	The number of users you follow.	Based on a sample of 1.2m users
18	5	#100,141,958

Figure 3.18.: Overview of Twitter Counter, Source http://twittercounter.com/pages/you, Accessed: 25 August 2015.

4. TwitterSuitcase

Many times after a Twitter event is finished, the majority of tweets are not accessible due to a lack of possibility to save those tweets and to view them again. Storing and analyzing the tweets during an event was successfully implemented by Thomas Altmann in his master's thesis [Altmann, 2014]. The process is divided into several services such as TweetCollector, TwitterStat[1] and TwitterWall[2].

TwitterSuitcase represents analyzing and visualizing wrapper functions of Twitter data collected by TweetCollector.

4.1. Why TwitterSuitcase?

Why TwitterSuitcase? Almost all applications about Twitter and tweets that can be found on the world wide web include the noun ”twitter” or ”tweet” in their name. This way, one can achieve a strong identification between the current application and Twitter, so the user has no doubts about the data source that will be used. During the initial discussions about the project and its implementation the same question was being repeatedly asked: ”What sort of information will be available to me when a Twitter event is finished?”. The response to this question would be a suitcase full of sorted and analyzed data.

[1] http://twitter.learninglab.tugraz.at/stat/, Accessed: 26 August 2015

[2] http://twitter.learninglab.tugraz.at/wall/, Accessed: 26 August 2015

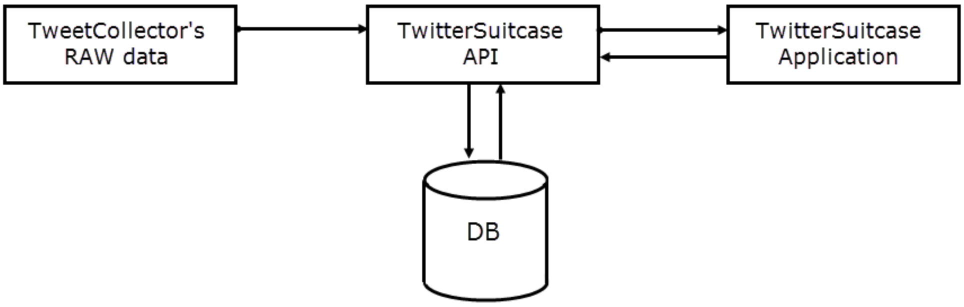


Figure 4.1.: Model overview.

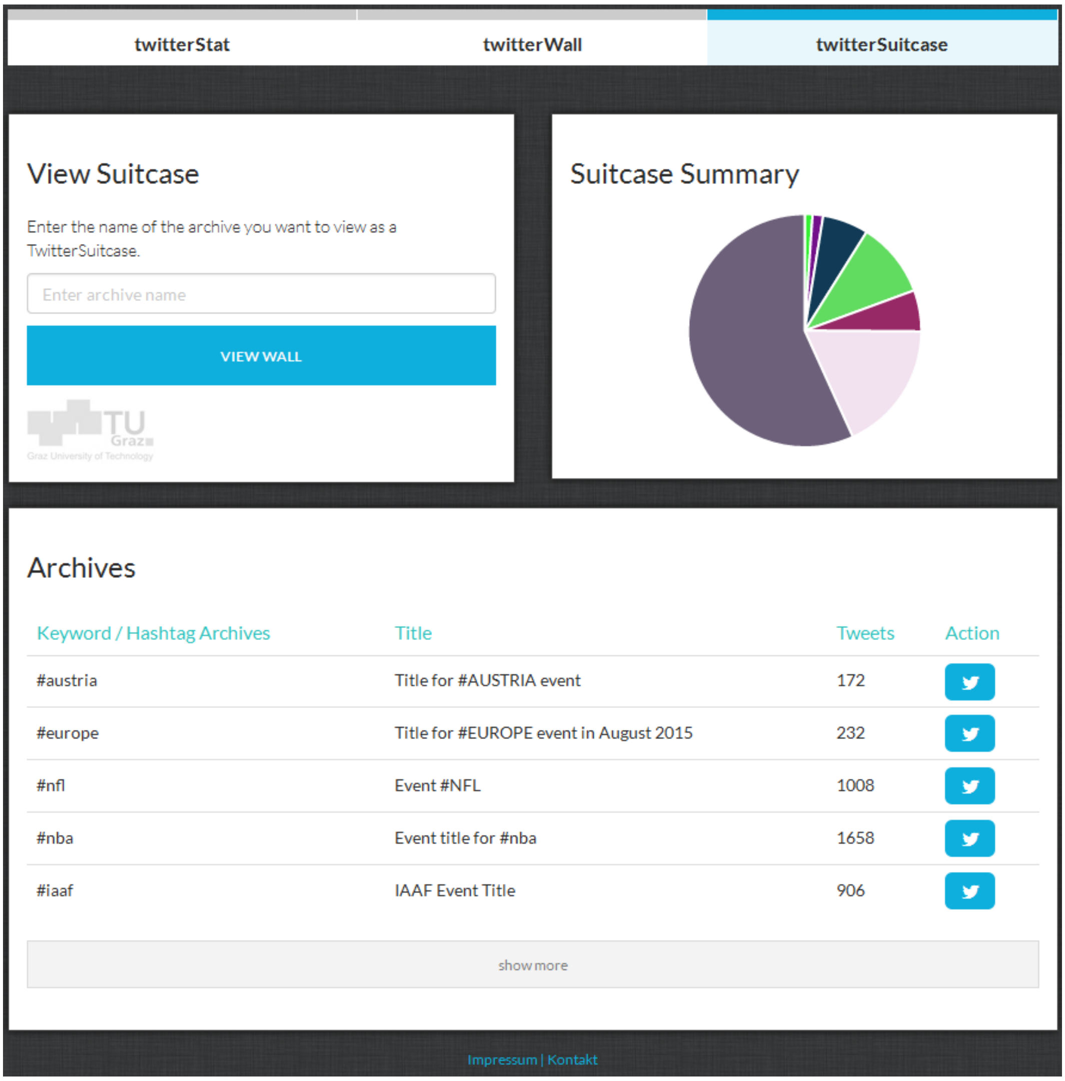
4.2. Concept

Figure 4.1 gives an overview of the structure and how each of the components are communicating to each other. TweetCollector represents the collection of tweets and builds a foundation for TwitterSuitcase. TwitterSuitcase API is a collection of functions and methods for analyzing, parsing and displaying tweets. The application's operations are bidirectional, depending on whether tweets are being initially created or just viewed by a random user.

The start page of TwitterSuitcase is divided into three boxes or categories, shown in Figure 4.2. The first box represents the navigation bar between TwitterStat, TwitterWall and TwitterSuitcase. The second box is divided into two columns where the first column is a wrapper in case one searches for hashtags. The second column offers a pie chart representation of the total amount of tweets grouped by a hashtag already processed by TwitterSuitcase. Finally, the last box presents TwitterSuitcase results for different hashtags. Those results include a keyword or hashtag, title of TwitterSuitcase and the total number of tweets for this event.

TwitterSuitcase for a single hashtag, shown in Figures 4.3 and 4.4, is divided into eight main categories and one additional category used as Wikipedia external links. Those eight main categories are:

- Top Users



View Suitcase

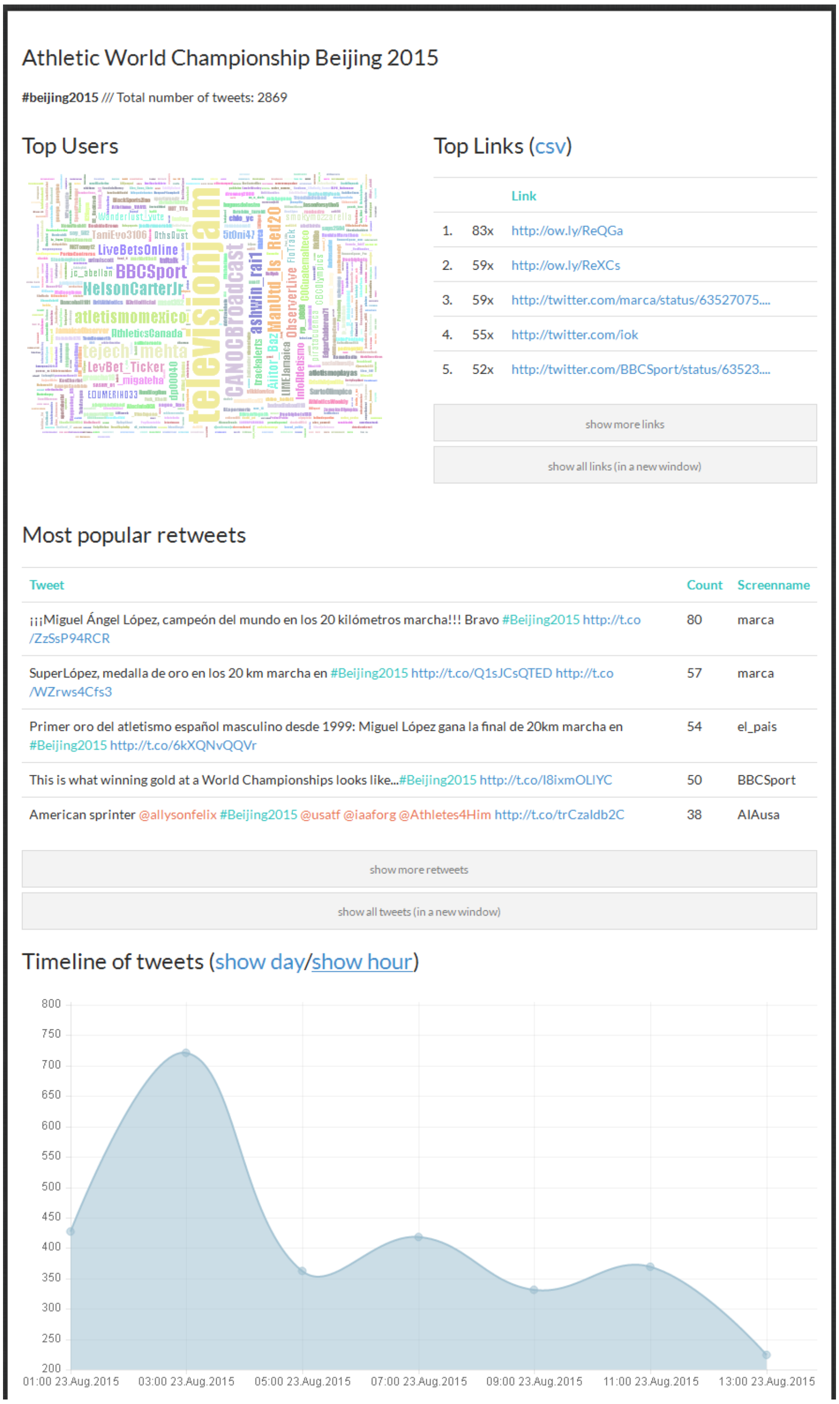
Enter the name of the archive you want to view as a TwitterSuitcase.

Suitcase Summary

Archives

Keyword / Hashtag Archives	Title	Tweets	Action
#austria	Title for #AUSTRIA event	172	🐦
#europe	Title for #EUROPE event in August 2015	232	🐦
#nfl	Event #NFL	1008	🐦
#nba	Event title for #nba	1658	🐦
#iaaf	IAAF Event Title	906	🐦

Figure 4.2.: TwitterSuitcase start page.



Athletic World Championship Beijing 2015

#beijing2015 /// Total number of tweets: 2869

Top Users

Top Links (csv)

		Link
1.	83x	http://ow.ly/ReQGa
2.	59x	http://ow.ly/ReXCs
3.	59x	http://twitter.com/marca/status/63527075....
4.	55x	http://twitter.com/iok
5.	52x	http://twitter.com/BBCSport/status/63523....

show more links

show all links (in a new window)

Most popular retweets

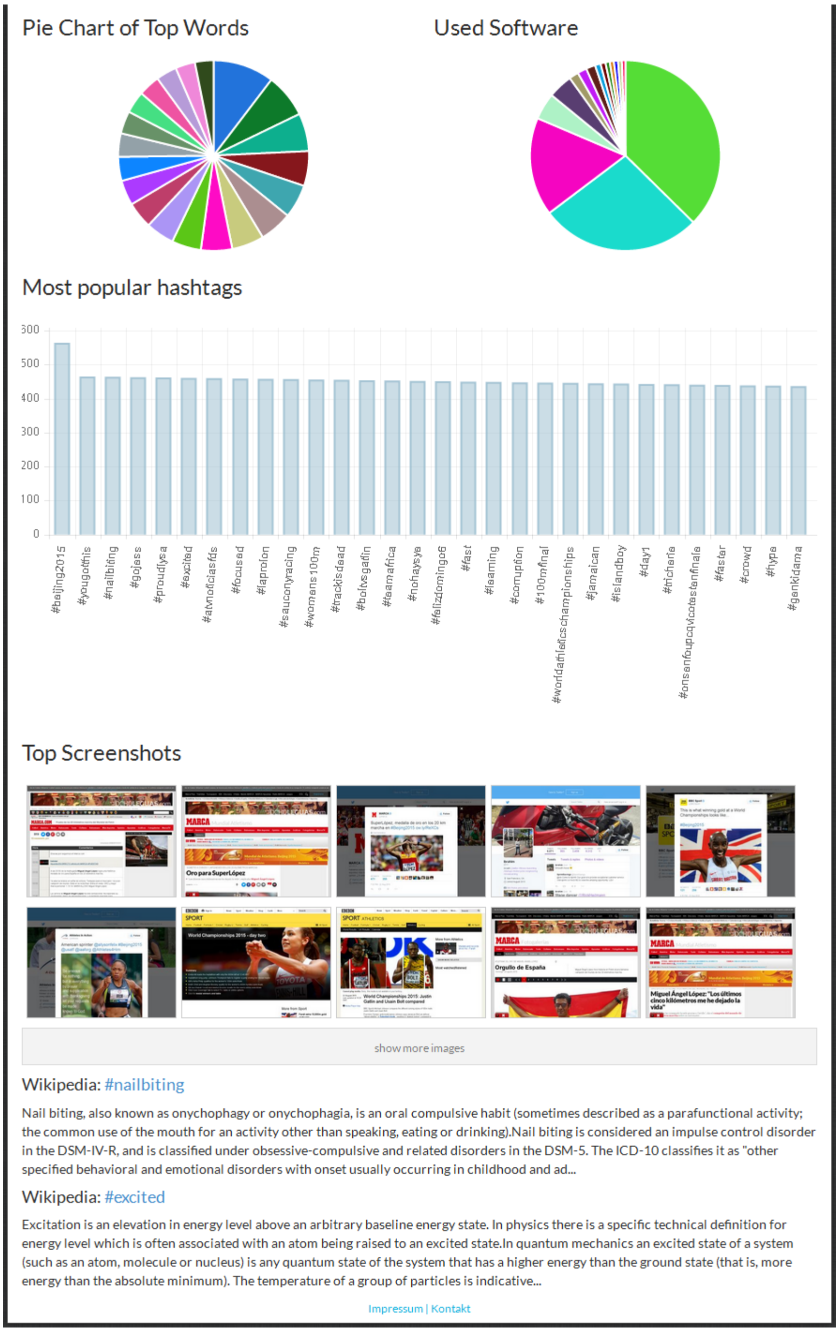
Tweet	Count	Screenname
¡¡¡Miguel Ángel López, campeón del mundo en los 20 kilómetros marcha!!! Bravo #Beijing2015 http://t.co/ZzSsP94RCR	80	marca
SuperLópez, medalla de oro en los 20 km marcha en #Beijing2015 http://t.co/Q1sJCsQTED http://t.co/WZrws4Cfs3	57	marca
Primer oro del atletismo español masculino desde 1999: Miguel López gana la final de 20km marcha en #Beijing2015 http://t.co/6kXQNvQQVr	54	el_pais
This is what winning gold at a World Championships looks like...#Beijing2015 http://t.co/l8ixmOLIYC	50	BBCSport
American sprinter @allysonfelix #Beijing2015 @usatf @iaaforg @Athletes4Him http://t.co/trCzaldb2C	38	AIAusa

show more retweets

show all tweets (in a new window)

Timeline of tweets (show day/show hour)

Figure 4.3.: TwitterSuitcase event page upper half.



Wikipedia: #nailbiting

Nail biting, also known as onychophagy or onychophagia, is an oral compulsive habit (sometimes described as a parafunctional activity; the common use of the mouth for an activity other than speaking, eating or drinking).Nail biting is considered an impulse control disorder in the DSM-IV-R, and is classified under obsessive-compulsive and related disorders in the DSM-5. The ICD-10 classifies it as "other specified behavioral and emotional disorders with onset usually occurring in childhood and ad...

Wikipedia: #excited

Excitation is an elevation in energy level above an arbitrary baseline energy state. In physics there is a specific technical definition for energy level which is often associated with an atom being raised to an excited state.In quantum mechanics an excited state of a system (such as an atom, molecule or nucleus) is any quantum state of the system that has a higher energy than the ground state (that is, more energy than the absolute minimum). The temperature of a group of particles is indicative...

Figure 4.4.: TwitterSuitcase event page bottom half.

- Top Links
- Most Popular Retweets
- Timeline Of Tweets
- Top Words
- Top Software
- Most Popular Hashtags
- Top Screenshots
- Wikipedia

Each of the listed elements will be briefly described in the following sections.

4.2.1. Top Users

This category represents the word cloud created from the list of all Twitter users who have written about the given hashtag (see Figure 4.5). The more tweets each user has, the bigger the representation of his username will be. The representation of the username is linked to the appropriate Twitter account of the user.

4.2.2. Top Links

Figure 4.6 presents the list of most popular HTTP links, including the number of occurrences within the event. The list of Top Links is restricted to SHOW_MAX_HTTP_LINKS_DETAIL (see section 4.5 for detailed information) which is set to twenty results by default. This means that clicking on ”Show more links” will display a maximum of twenty HTTP links grouped in five-element blocks. There is also an option to export all displayed links in external file in Comma Separated Value (CSV) format. The button ”show all links” will open all available links in a new window. This is described in section 4.2.11.

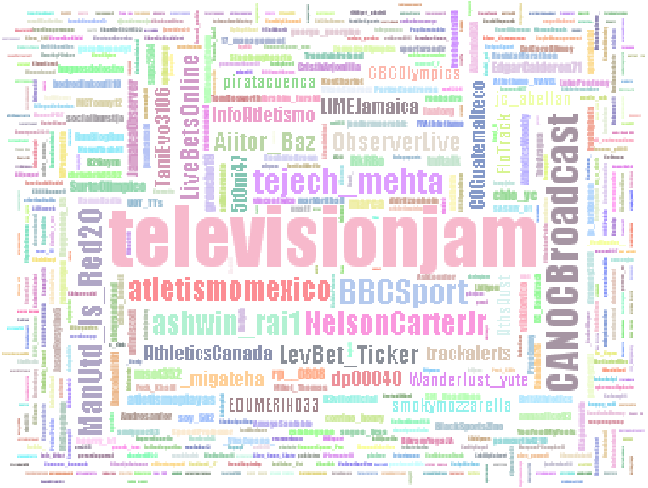
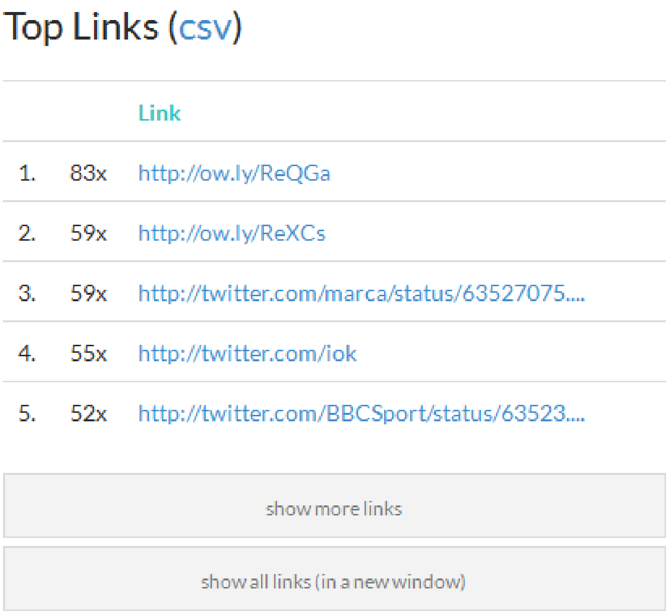


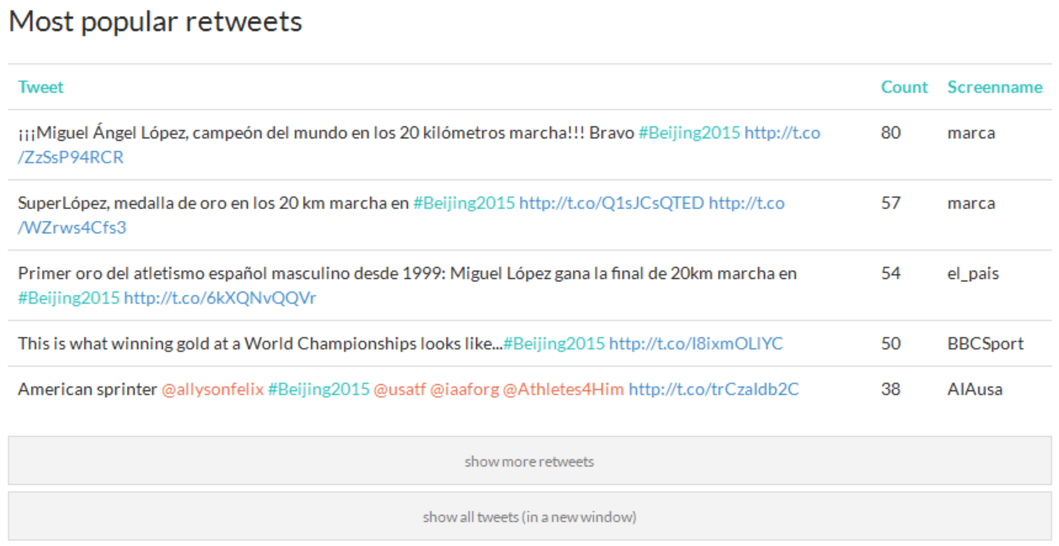
Figure 4.5.: Top Users word cloud.



Top Links (csv)

		Link
1.	83x	http://ow.ly/ReQGa
2.	59x	http://ow.ly/ReXCs
3.	59x	http://twitter.com/marca/status/63527075....
4.	55x	http://twitter.com/iok
5.	52x	http://twitter.com/BBCSport/status/63523....

Figure 4.6.: List of Top Links.



Most popular retweets

Tweet	Count	Screenname
¡¡¡Miguel Ángel López, campeón del mundo en los 20 kilómetros marcha!!! Bravo #Beijing2015 http://t.co/ZzSsP94RCR	80	marca
SuperLópez, medalla de oro en los 20 km marcha en #Beijing2015 http://t.co/Q1sJCsQTED http://t.co/WZrws4Cfs3	57	marca
Primer oro del atletismo español masculino desde 1999: Miguel López gana la final de 20km marcha en #Beijing2015 http://t.co/6kXQNvQQVr	54	el_pais
This is what winning gold at a World Championships looks like...#Beijing2015 http://t.co/l8ixmOLIYC	50	BBCSport
American sprinter @allysonfelix #Beijing2015 @usatf @iaaforg @Athletes4Him http://t.co/trCzaldb2C	38	AIAusa

Figure 4.7.: List of most popular retweets.

4.2.3. Most Popular Retweets

The list of most popular retweets, shown in Figure 4.7, correlates to the list of tweet appearances within the event. The list on the event page is restricted to SHOW_MAX_TWEETS_DETAIL (see section 4.5 for detailed information) and the default value is set to twenty tweets. Hence, by clicking on ”show more retweets”, the tweets will be displayed in five-element blocks until the maximum value is reached. All available tweets can be seen by clicking on ”show all tweets” button. This will trigger opening a new window displaying all tweets. This is described in section 4.2.10.

4.2.4. Timeline of Tweets

A chart is designed to always show at least two different timeline values. Those values depend on whether the event took place over several hours, days, months or even years. All of those values are grouped and the largest groups having more than two elements are displayed. The groups are

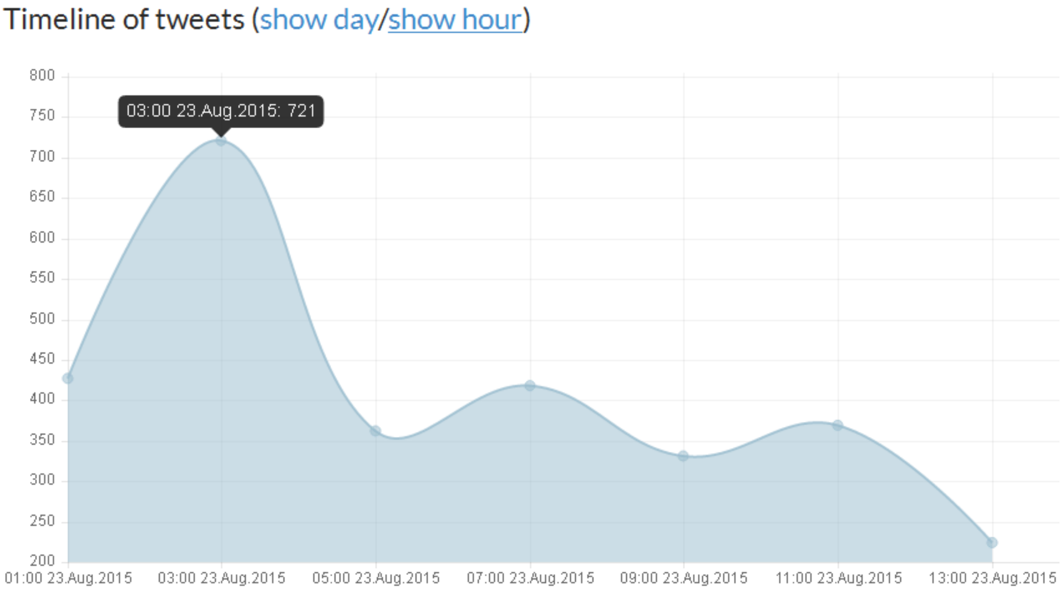


Figure 4.8.: Timeline of tweets.

created in accordance with to the following sort rule:

Year >Month >Month >Day.

In Figure 4.8 the charts show the timeline for ”day” and ”hours”.

4.2.5. Top Words

During the analysis and data processing, all words appearing within the event are counted. The top twenty words are displayed in a pie chart which is demonstrated in Figure 4.9.

4.2.6. Top Software

Top software represents the top twenty software or applications used to send a tweet (see Figure 4.10).

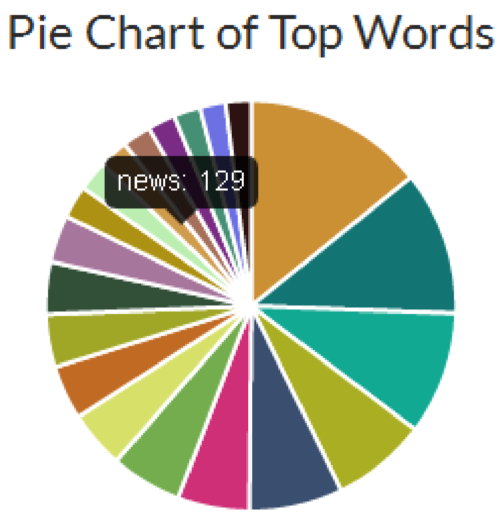


Figure 4.9.: Top words.

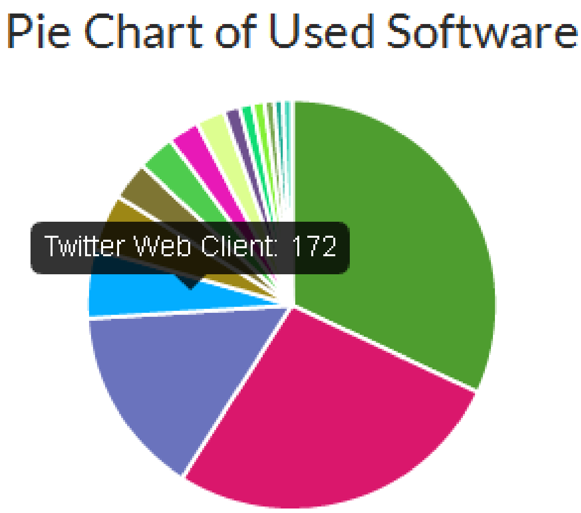


Figure 4.10.: Top software or applications.

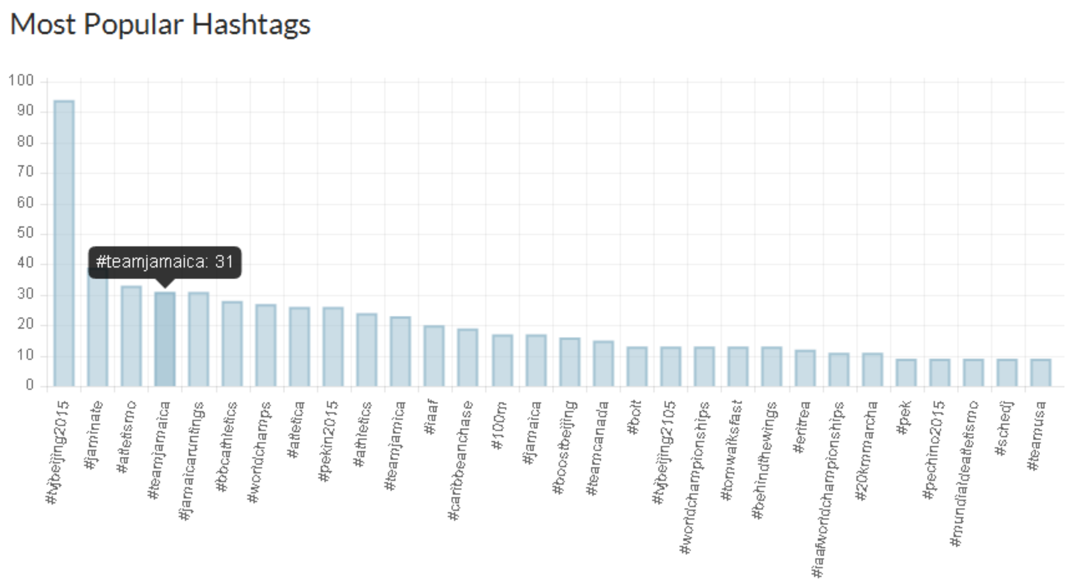


Figure 4.11.: Top used hashtags.

4.2.7. Most Popular Hashtags

The chart in Figure 4.11 shows the number of occurences for the top thirty hashtags. Although all hashtags have been counted for their occurrence only the most frequent ones are being displayed.

4.2.8. Top Screenshots

Top Screenshots are snapshots of the most popular HTTP links displayed in section 4.2.2. The principle of displaying the screenshots is the same as for the Top Links and Most Popular Retweets where five-element blocks are displayed until a certain threshold is reached. When clicking on one of the snapshots the screenshot will be shown including one of the tweets in which the HTTP links used for the snapshot appears (see Figure 4.13). Google API

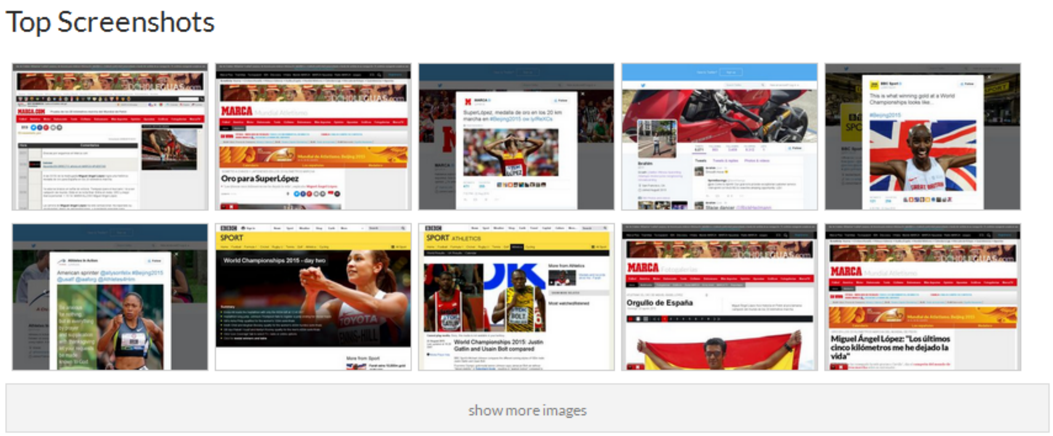


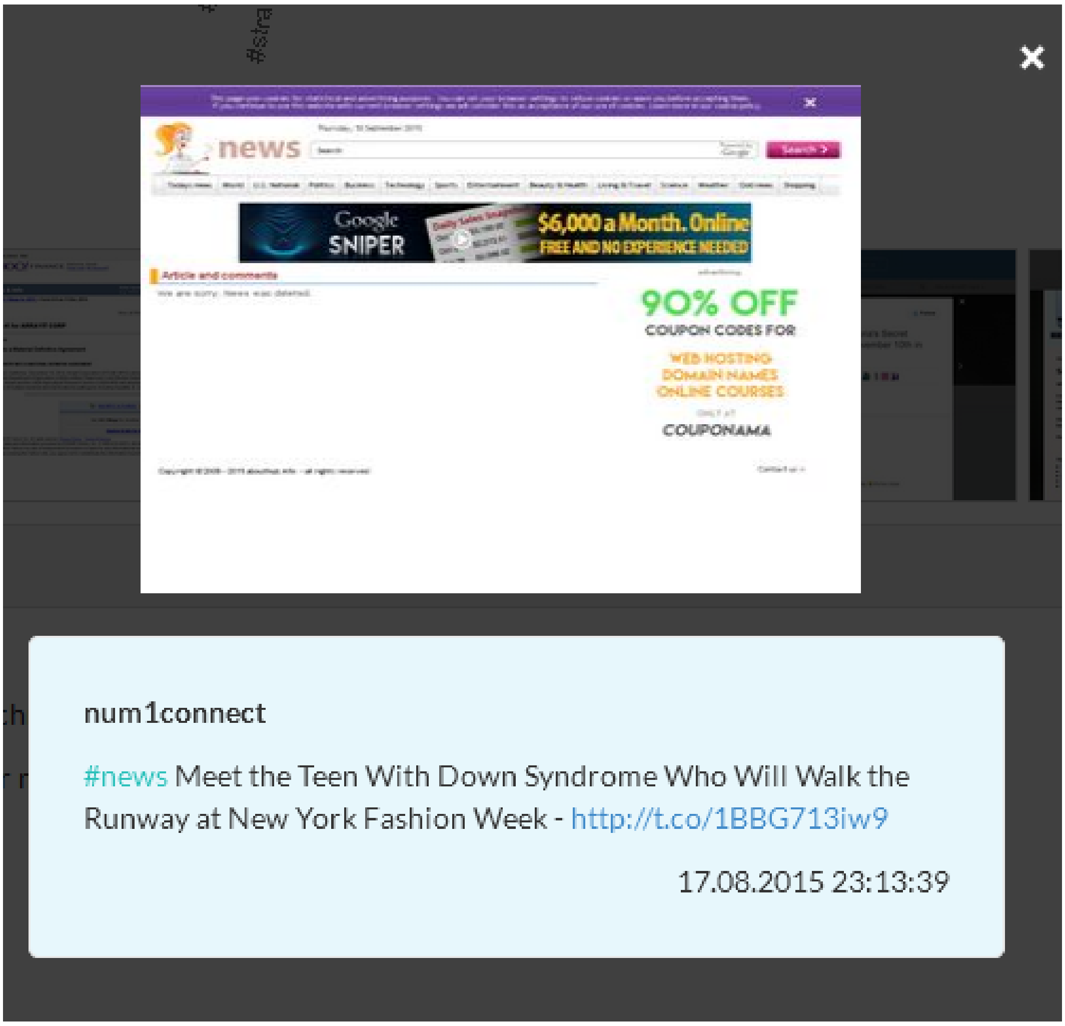
Figure 4.12.: Top screenshots.

PageSpeed[3] Insights have been used for the creation of the snapshot.

4.2.9. Wikipedia

Top five of the most popular hashtags (see section 4.2.7) are used as triggers for the Wikipedia search API. If the search engine retrieves a meaningful results, then the first 500 characters of the article are displayed, as shown in Figure 4.14. The title of each hashtag is an HTML reference to an appropriate Wikipedia site. The results are classified into partial and full results depending on what sort of information they are retrieving. Partial results are those results that provide a list of possible referrers or article recommendations, but do not have the main Wikipedia article. Full results are results leading to the main Wikipedia article handling the given hashtag.

[3] https://developers.google.com/speed/pagespeed/insights/, Accessed: 26 August 2015

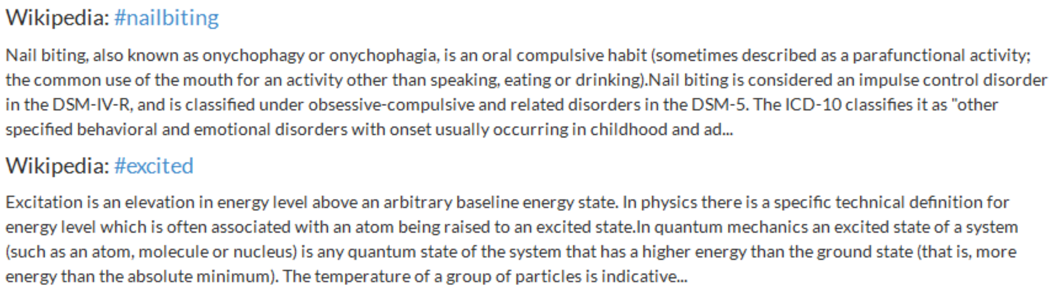


num1connect

#news Meet the Teen With Down Syndrome Who Will Walk the Runway at New York Fashion Week - http://t.co/1BBG713iw9

17.08.2015 23:13:39

Figure 4.13.: Top Screenshot detail.



Wikipedia: #nailbiting

Nail biting, also known as onychophagy or onychophagia, is an oral compulsive habit (sometimes described as a parafunctional activity; the common use of the mouth for an activity other than speaking, eating or drinking).Nail biting is considered an impulse control disorder in the DSM-IV-R, and is classified under obsessive-compulsive and related disorders in the DSM-5. The ICD-10 classifies it as "other specified behavioral and emotional disorders with onset usually occurring in childhood and ad...

Wikipedia: #excited

Excitation is an elevation in energy level above an arbitrary baseline energy state. In physics there is a specific technical definition for energy level which is often associated with an atom being raised to an excited state.In quantum mechanics an excited state of a system (such as an atom, molecule or nucleus) is any quantum state of the system that has a higher energy than the ground state (that is, more energy than the absolute minimum). The temperature of a group of particles is indicative...

Figure 4.14.: Wikipedia results.

4.2.10. Show All Tweets (in a new window)

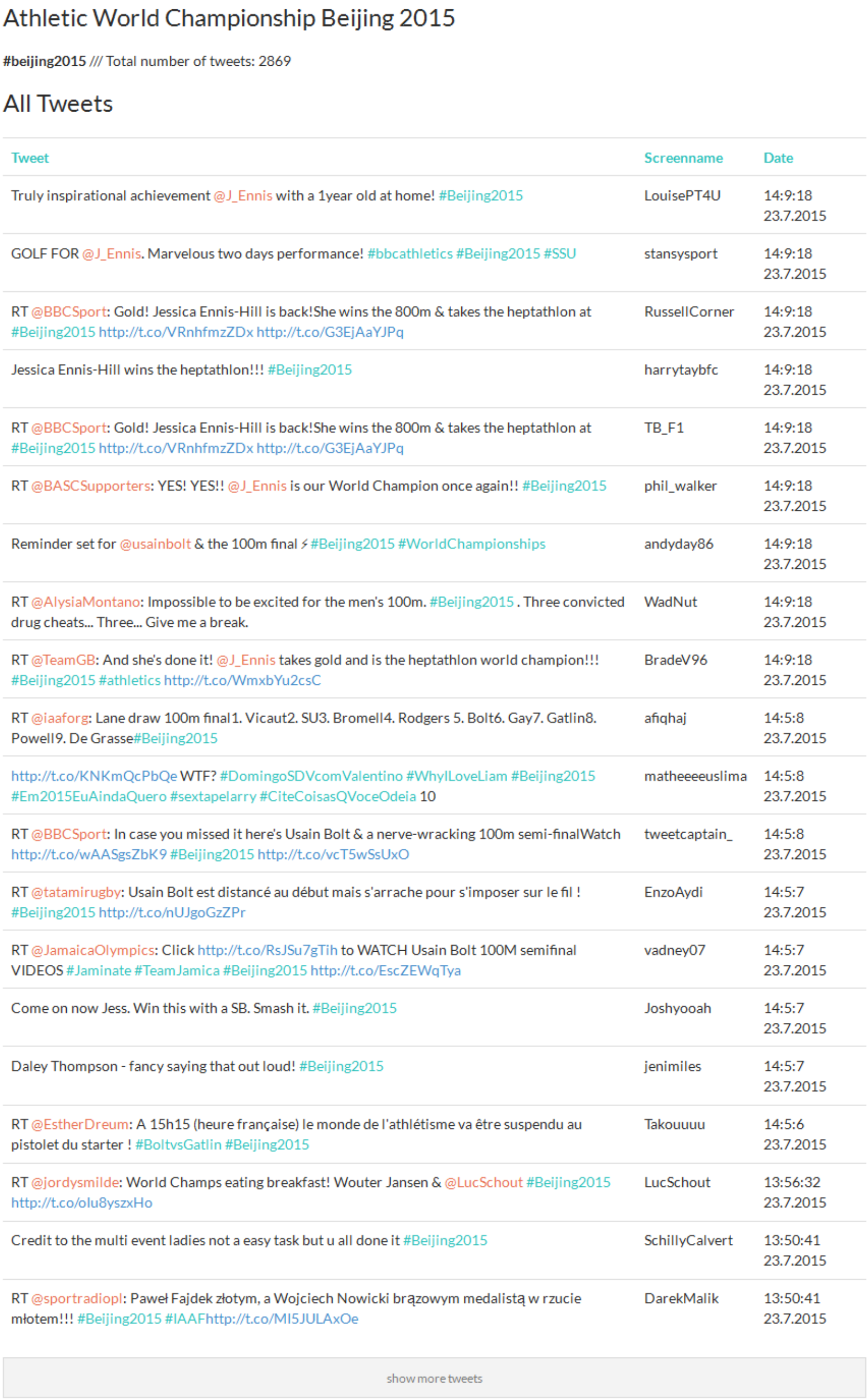
In the case that a user wants to see all tweets collected during a specific event, button ”show all tweets” showed in Figure 4.3 needs to be clicked on. By clicking this button a new window will be opened with the list of all collected tweets (see Figure 4.15). Since the observed event could be popular and the list may be very long and include thousands of tweets, the length of the list needs to be restricted to a certain threshold. This threshold is set to twenty results by default and its usage is explained in section called Requirements and Configuration 4.5.

4.2.11. Show All Links (in a new window)

The same principle, as described in previous chapter is applied to the list of all HTTP links. The only difference is the possibility to extract all links into a CSV data format and to use them in an external program or applications, such as Microsoft Excel. The Figure 4.16 gives an overview of such list.

4.3. Implementation

This section explains how each of the components works and what kind of functionality it gives.



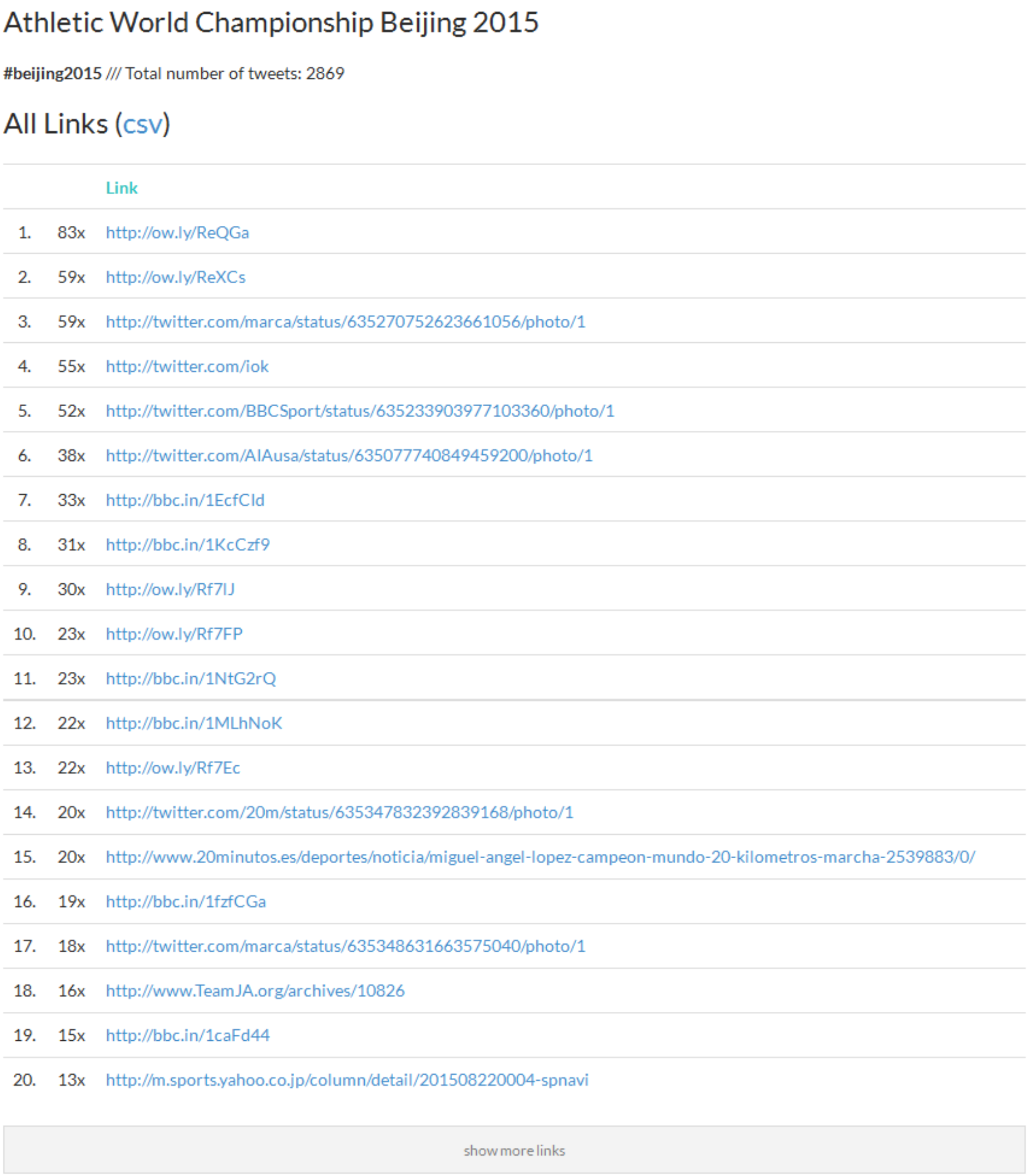
Athletic World Championship Beijing 2015

#beijing2015 /// Total number of tweets: 2869

All Tweets

Tweet	Screenname	Date
Truly inspirational achievement @J_Ennis with a 1year old at home! #Beijing2015	LouisePT4U	14:9:18 23.7.2015
GOLF FOR @J_Ennis. Marvelous two days performance! #bbcathletics #Beijing2015 #SSU	stansysport	14:9:18 23.7.2015
RT @BBCSport: Gold! Jessica Ennis-Hill is back!She wins the 800m & takes the heptathlon at #Beijing2015 http://t.co/VRnhfmzZDx http://t.co/G3EjAaYJPq	RussellCorner	14:9:18 23.7.2015
Jessica Ennis-Hill wins the heptathlon!!! #Beijing2015	harrytaybfc	14:9:18 23.7.2015
RT @BBCSport: Gold! Jessica Ennis-Hill is back!She wins the 800m & takes the heptathlon at #Beijing2015 http://t.co/VRnhfmzZDx http://t.co/G3EjAaYJPq	TB_F1	14:9:18 23.7.2015
RT @BASCSupporters: YES! YES!! @J_Ennis is our World Champion once again!! #Beijing2015	phil_walker	14:9:18 23.7.2015
Reminder set for @usainbolt & the 100m final ⚡ #Beijing2015 #WorldChampionships	andyday86	14:9:18 23.7.2015
RT @AlysiaMontano: Impossible to be excited for the men's 100m. #Beijing2015 . Three convicted drug cheats... Three... Give me a break.	WadNut	14:9:18 23.7.2015
RT @TeamGB: And she's done it! @J_Ennis takes gold and is the heptathlon world champion!!! #Beijing2015 #athletics http://t.co/WmxbYu2csC	BradeV96	14:9:18 23.7.2015
RT @iaaforg: Lane draw 100m final1. Vicaut2. SU3. Bromell4. Rodgers 5. Bolt6. Gay7. Gatlin8. Powell9. De Grasse#Beijing2015	afiqhaj	14:5:8 23.7.2015
http://t.co/KNKmQcPbQe WTF? #DomingoSDVcomValentino #WhyILoveLiam #Beijing2015 #Em2015EuAindaQuero #sextapelarry #CiteCoisasQVoceOdeia 10	matheeeeuslima	14:5:8 23.7.2015
RT @BBCSport: In case you missed it here's Usain Bolt & a nerve-wracking 100m semi-finalWatch http://t.co/wAASgsZbK9 #Beijing2015 http://t.co/vcT5wSsUxO	tweetcaptain_	14:5:8 23.7.2015
RT @tatamirugby: Usain Bolt est distancé au début mais s'arrache pour s'imposer sur le fil ! #Beijing2015 http://t.co/nUJgoGzZPr	EnzoAydi	14:5:7 23.7.2015
RT @JamaicaOlympics: Click http://t.co/RsJSu7gTih to WATCH Usain Bolt 100M semifinal VIDEOS #Jaminate #TeamJamica #Beijing2015 http://t.co/EscZEWqTya	vadney07	14:5:7 23.7.2015
Come on now Jess. Win this with a SB. Smash it. #Beijing2015	Joshyooah	14:5:7 23.7.2015
Daley Thompson - fancy saying that out loud! #Beijing2015	jenimiles	14:5:7 23.7.2015
RT @EstherDreum: A 15h15 (heure française) le monde de l'athlétisme va être suspendu au pistolet du starter ! #BoltvsGatlin #Beijing2015	Takouuuu	14:5:6 23.7.2015
RT @jordysmilde: World Champs eating breakfast! Wouter Jansen & @LucSchout #Beijing2015 http://t.co/olu8yszxHo	LucSchout	13:56:32 23.7.2015
Credit to the multi event ladies not a easy task but u all done it #Beijing2015	SchillyCalvert	13:50:41 23.7.2015
RT @sportradiopl: Paweł Fajdek złotym, a Wojciech Nowicki brązowym medalistą w rzucie młotem!!! #Beijing2015 #IAAFhttp://t.co/MI5JULAxOe	DarekMalik	13:50:41 23.7.2015

Figure 4.15.: Show all tweets within the event.



Athletic World Championship Beijing 2015

#beijing2015 /// Total number of tweets: 2869

All Links (csv)

		Link
1.	83x	http://ow.ly/ReQGa
2.	59x	http://ow.ly/ReXCs
3.	59x	http://twitter.com/marca/status/635270752623661056/photo/1
4.	55x	http://twitter.com/iok
5.	52x	http://twitter.com/BBCSport/status/635233903977103360/photo/1
6.	38x	http://twitter.com/AIAusa/status/635077740849459200/photo/1
7.	33x	http://bbc.in/1EcfCId
8.	31x	http://bbc.in/1KcCzf9
9.	30x	http://ow.ly/Rf7IJ
10.	23x	http://ow.ly/Rf7FP
11.	23x	http://bbc.in/1NtG2rQ
12.	22x	http://bbc.in/1MLhNoK
13.	22x	http://ow.ly/Rf7Ec
14.	20x	http://twitter.com/20m/status/635347832392839168/photo/1
15.	20x	http://www.20minutos.es/deportes/noticia/miguel-angel-lopez-campeon-mundo-20-kilometros-marcha-2539883/0/
16.	19x	http://bbc.in/1fzfCGa
17.	18x	http://twitter.com/marca/status/635348631663575040/photo/1
18.	16x	http://www.TeamJA.org/archives/10826
19.	15x	http://bbc.in/1caFd44
20.	13x	http://m.sports.yahoo.co.jp/column/detail/201508220004-spnavi

Figure 4.16.: Show all HTTP links within the event.

4.3.1. TwitterSuitcase creating and deleting

There are two basic features when considering TwitterSuitcase, namely creating and deleting.

When creating a new TwitterSuitcase, the main requirement is that the collection of tweets is already established by TweetCollector. This requirement is basically always fulfilled, since the web interface is showing only the list of events for which the tweets have been collected. Nevertheless, if a user has gained basic information about this service and its principle of functionality, then the web interface can be bypassed by accessing directly to the API (explained in TwitterSuitcase API 4.3.2) using random generated data. This kind of backdoor approach could seriously influence the integrity of the system, due to the security data-model checks which run after the initial create request has been started, and is therefore not allowed.

Just like creating, deleting is also possible only in case TwitterSuitcase was already generated including security and data-model checks in the background.

4.3.2. TwitterSuitcase API

TwitterSuitcase provides a single API which covers all of its functionality. This API is implemented in PHP and stored as a ”data.php” file on server. It retrieves data, depending on the requested parameters, in JSON data format. The API supports following actions:

- *Create TwitterSuitcase*
  Creates TwitterSuitcase out of a collection of tweets.
- *Delete TwitterSuitcase*
  Deletes TwitterSuitcase for specific hashtag but does not remove data collected by TweetCollector.
- *Display single TwitterSuitcase*
  Display all information about a single hashtag.
- *Display all TwitterSuitcase's*
  Display all processed hashtags without detailed information.

- *Display all HTTP's from single TwitterSuitcase*
  Show all HTTP links for a specific hashtag.
- *Display all Tweets from single TwitterSuitcase*
  Display all tweets for a specific hashtag.
- *Display CSV structure for HTTP's from the TwitterSuitcase*
  Display CSV structured HTTP links for a specific hashtag
- *Display all saved events*
  Display all archives already collected with Tweetcollector.

The code snippet in Listing 4.1 is an example of an API call in order to retrieve a list of all created TwitterSuitcases from the database. Such list is used on the start page of the TwitterSuitcase when displaying results for all hashtags. Listing 4.2 shows the API response in the form of the JSON data format. This piece of code returns the information (id, hashtag, title of the TwitterSuitcase and total number of tweets) about previously created TwitterSuitcases for hashtags #iaaf, #beijing2015 and #football. Although the retrieved number of results is equivalent to the total number of TwitterSuitcases, the demonstrated response code has been reduced for the purpose of readability.

```
if (isset($_GET[SUITCASE]) && $_GET[SUITCASE] == "all") {
        echo json_encode($tp->wbGetAllArchivesFromSuitcase());
}
```

Listing 4.1: API call to show all TwitterSuitcases

```
[
 {"id":"79"
 ,"keyword":"#iaaf"
 ,"title":"IAAF Event Title "
 ,"count":"906"},
 {"id":"80"
 ,"keyword":"#beijing2015"
 ,"title":"Athletic World Championship Beijing 2015"
 ,"count":"2869"},
 {"id":"81"
 ,"keyword":"#football"
 ,"title":"Football TwitterEvent 2015"
 ,"count":"8991"}
]
```

Listing 4.2: JSON Response to API call

The following URL call
http://DOMAIN/suitcase/api/data.php?id=76&action=show
shows the possibility of exploiting data which was created by TwitterSuitcase over an API call for embedding results within some external interfaces for the purpose of further data processing. An appropriate server response is shown in Listing 4.3 for a #europe hashtag. The response includes all information, such as top users, top links, most popular retweets, timeline of tweets, pie charts of top words and used software, hashtag popularity overview, top screenshots and possible Wikipedia articles in JSON data format. Listing 4.3 presents a short version of the server response.

```
{
"archive":{
  "archive_id":"76","title":"Europa title","keyword":"\#europe"
      ,"max_count":"232","count":"232","interval_begin":"August
      5th 2015 21:42:38","interval_end":"August 5th 2015
      22:44:12"},
"https":{
  "https://www.flickr.com/photos/130169970l286149/": {"
      contenttype":"unknown","total\_tweets":"3","id":["
      629029200356352001","629025175024103424","
      629023163872071680"]}},
"users":{
  "Pr45H8tZi":{"total\_tweets":10,"user\_id":"403296700","
      629029514920726528":{"id":"629029514920726528","screen\
      _name":"Pr45H8tZi","user\_id":"403296700","source":"
      Instagram<\/a>","text":"Boat Cruise in the Paris canal. \#
      paris \#France \#Europe \#holidays \#fun \#cityoflove \#
      beautiful\u2026 https:\/\/t.co\/dj7oEu8hpI","created\_at":
      "1438807297","in\_reply\_to\_screen\_name":"","in\_reply\
      _to\_status\_id":"0","in\_reply\_to\_user\_id":"0","
      coordinates\_type":"","coordinates\_lat":"0","coordinates\
      _long":"0"}}},
"tweets":{
  "The way to stop these tragedies for good? The creation of
      safe and legal ways for people to seek asylum or migrate
      to #Europe":{"screen_name":"MSF_Sea","total_tweets":6,"id"
      :["629024993913999360","629024878579068928","
      629024600198918145","629023434522132480","
      629023278154317824","629023235988963333"]}},
"words":{
  "to":{"occurred":109},"the":{"occurred":103},"rt":{"occurred"
      :83}},
"hashtags":{
  "labels":["#europe","#travel","#france","#paris","#fun",...],
      "datasets":[{"data":[212,29,15,12,12,...]}]},
"software":{
  "Twitter Web Client":{"occurred":37},"Instagram":{"occurred"
      :36},"Twitter for Android":{"occurred":30}},
"timeline":{
  "day":{"labels":["26.Sep.2015","27.Sep.2015"],"datasets":[{"
      data":["4900","4100"]]}}},
"screenshot":[
  {"mime_type":"image\/jpeg","id":{"id":"629029200356352001","
      text":"Truth...https:\/\/t.co\/Pkkf7Rpcxo#news#BBC#CNN#
      worldnews#foxnews#world#USANews#Europe http:\/\/t.co\/
      tFtJYp74B1","screen_name":"j_marcello","created_at":"
      05.08.2015 22:40:22","user_id":"3007680186"},"image":"\/9j
      \/4AAQSkZJRgABAQAAAQABAAD\/2wBDAAYEBQYFBAYGBQYHBwYIC"}]
}
```

Listing 4.3: API call response for #europe hashtag

4.4. User Interface

The HTML files used by an average user are split into three Graphical User Interfaces (GUI):

- *index.html* : Start page
- *suitcase.html* : Event page
- *detail.html* : Event detail page

In relation to those three GUIs there are also three JavaScript files whose purpose is to ensure the functionality of each GUI:

- index.js
- suitcase.js
- detail.js

An example of the start page or "index.html" is shown in Figure 4.2. From a random user's point of view, this is the start and the first thing one can see is the list of available TwitterSuitcases created from the data collection of the TweetCollector. If the list is too large, one can search for a specific keyword or hashtag by entering the search term in the search box and clicking the search button. If the entered value does not correspond to the parsed data, the user will receive a message that the searched value does not exist as a TwitterSuitcase. Notice that a collected set of data could be provided by TweetCollector although it has still not been processed in the TwitterSuitcase application.

Once the keyword is found, the user will automatically be redirected to Event page or ”suitcase.html” where all details from within this event can be found (see Figures 4.3 and 4.4).

If one wants to see all tweets or HTTP links within the event, he or she will be redirected to an event detail page or ”detail.html” (see Figures 4.15 or 4.16) where the full list of tweets or HTTP links is presented.

4.5. Requirements and Configuration

This section provides more detailed information about the requirements and configurations of TwitterSuitcase.

4.5.1. Dependencies

The usage of TwitterSuitcase requires the existence of a database collection (previously collected by TweetCollector). Furthermore, there is also a necessity for PHP version 5.6 or higher and MySQL database should be version 5.6 or higher.

4.5.2. Libraries

The following libraries are needed to ensure proper functionality: jQuery, Chart, awesomeCloud and Bootstrap.

jQuery

The used version of jQuery[4] in this thesis is 2.0.3. Since jQuery supports backward compatibility, it is possible to use an even higher version than the actual one.

Chart

Chart[5] is an Open Source JavaScript-based library for development and presentation of different types of charts by using HTML5[6]. The used version of this library is 1.0.2.

awesomeCloud

awesomeCloud is a jQuery-based plugin used for creation of word clouds in various shapes. The used version is 0.2

Bootstrap

Bootstrap 3, a Front-End framework which is used in this master's thesis is collection of tools for development of web applications and websites.

4.5.3. Configuration

The configuration of the project is divided into Front-End (client) and Back-End (server) configuration.

[4] https://jquery.com/, Accessed: 26 August 2015

[5] http://www.chartjs.org/, Accessed: 26 August 2015

[6] http://www.w3schools.com/html/, Accessed: 26 August 2015

Config.php

Config.php is used in applications for a configuration of the server-side. For the configuration to be successful, the following parameters needs to be adjusted (or controlled):

- *dbname* : Set the database name.
- *php_time_limit* : Set the PHP timeout.
- *php_mem_limit* : Set the memory limit.

The following list refers to the maximum number of selected data from the database:

- *LIMIT_TOP_LINKS* : HTTP links.
- *LIMIT_TOP_USERS* : Twitter users.
- *LIMIT_POPULAR_TWEETS* : Retweets.
- *LIMIT_TOP_WORDS* : Words.
- *LIMIT_TOP_HASHTAGS* : Hashtags.
- *LIMIT_TOP_SOFTWARE* : Software or Applications.

Config.js

In contrast to server-side, Config.js is used for defining client-side display options which include display routes and maximum display values.

Display routes are basically URL locations for:

- TwitterStat
- TwitterWall
- TwitterSuitcase

Display values on the client side defined in the exact same way that is outlined in section 4.5.3.

5. Evaluation

The previous chapter of this paper described the tool created for the analysis and visualization of the tweets. The next step, which will be realized in this chapter, is to evaluate the real-time tweets for a specific events in order to verify the functionality of the presented tool. The following two hashtags have been chosen for building up the test bundles:

- Use Case #news
- Use Case #emoocs2014

The following two sections contains detailed explanations of the results for each of the used hashtags.

5.1. Use Case #news

5.1.1. Why #news

News is a collection of information about some current or recent events which have happened somewhere in the world. The news is related to different kind of media like printing, broadcasting, word of mouth and electronic communication. This is the reason why the hashtag #news can be considered as very popular and the most commonly used hashtag within tweets.

5.1.2. Using TwitterSuitcase with #news

The first task was to provide the TweetCollector with appropriate hashtags and to determine its run time (meaning start and end date). For this purpose, the hashtags were started on 5th of August, 2015 at 22:00, because the hashtag is very popular, it was stopped two hours later on the same day (meaning before midnight) in order to avoid a lot of unwanted tweets. The same approach was repeated on 15th an 17th of August, 2015. The reason why this fractional approach was used lies in the fact that the collection of gathered tweets should stretch across multiple days while the amount of tweets should remain under 10.000. The functionality and the usage of the TweetCollector is described by Thomas Altmann in his master's thesis [Altmann, 2014]. The second step should initiate the creation of the TwitterSuitcase for the mentioned hashtag. This is done by using the administration interface as can be seen in Figure 5.1. The duration of the creation process, which includes data analysis and structure setup for 3740 tweets, was 165 seconds. After the TwitterSuitcase is created, all the information is being saved into the database.

On the start page of the TwitterSuitcase (see Figure 4.2) the user can choose the keyword #news for further analysis. After clicking on this keyword, the event page will be displayed (see Figures 4.3 and 4.4).

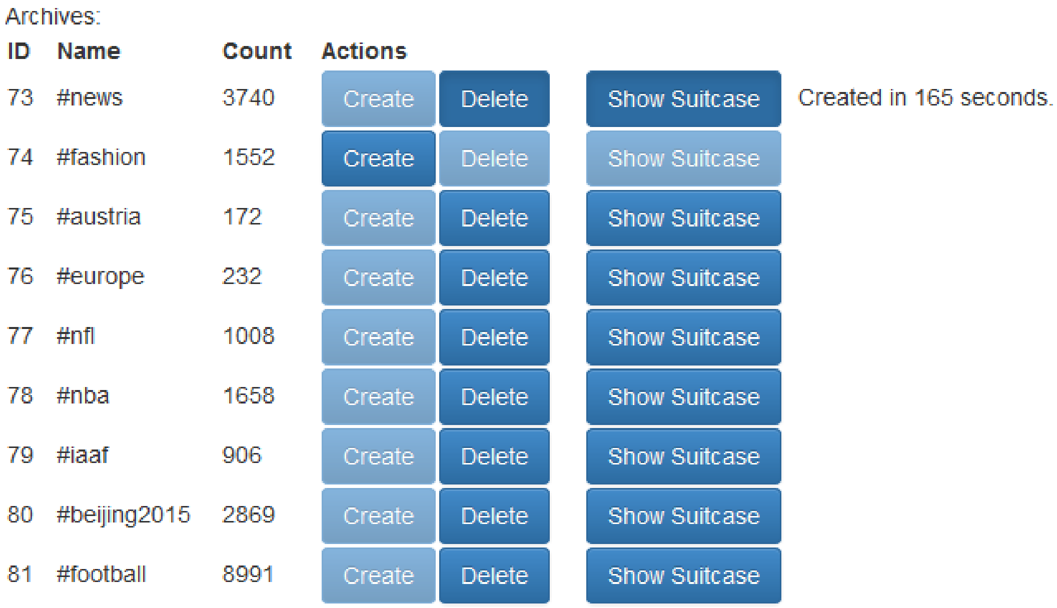
One can see the title of the event "Event #NEWS on August 2015", the used hashtag "#news", date and time of the first and last collected tweet and the total amount of tweets "3740" used in TwitterSuitcase.

From the word cloud of top users we can see that the largest number of tweets (double-digits) came from the following accounts: Setting4Success, mohammeddki8851, Beduac, til_now, FatenXerox, CollectedN, OlteenRazvan, news24husa, ru_newsmix, news24heng, berserk_news, news_in, Tukang_Update, 2chfind, bestnewrapmusic and so on.

In Figure 5.2, the word cloud can be inspected for further users.

The list of top links has created the following results (including the number of appearances):

- http://tinyurl.com/qjgf87m (89)



Archives:

ID	Name	Count	Actions			
73	#news	3740	Create	Delete	Show Suitcase	Created in 165 seconds.
74	#fashion	1552	Create	Delete	Show Suitcase	
75	#austria	172	Create	Delete	Show Suitcase	
76	#europe	232	Create	Delete	Show Suitcase	
77	#nfl	1008	Create	Delete	Show Suitcase	
78	#nba	1658	Create	Delete	Show Suitcase	
79	#iaaf	906	Create	Delete	Show Suitcase	
80	#beijing2015	2869	Create	Delete	Show Suitcase	
81	#football	8991	Create	Delete	Show Suitcase	

Figure 5.1.: Overview of the creation process.

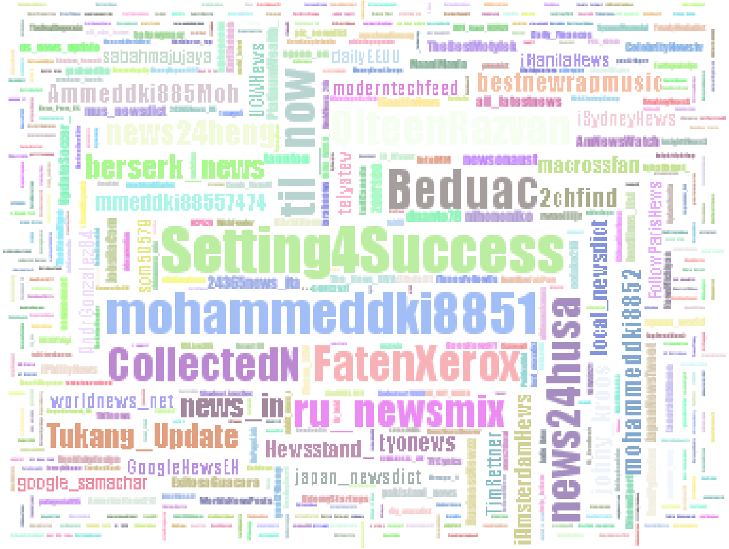


Figure 5.2.: Results for #news word cloud.

- http://otc.ninja/r/?s=ARYC&t=2a9d317a16872d2437703dd93c87cb77 (18)
- http://goo.gl/N2ZS5W (13)
- http://twitter.com/WikiNut_Blue/status/632591189984833536/photo/1 (13)
- http://bit.ly/1TQ7fZV (12)
- http://bit.ly/1gREQAJ (12)
- http://bit.ly/1gRER7z (12)
- http://twitter.com/VSangelslove/status/632579233798332416/photo/1 (12)
- http://bit.ly/1TQ7eVH (12)
- http://bit.ly/1gRETwp (12)
- https://twibble.io (11)
- http://MyShareInvest.com (9)
- http://ch.beduac.de (9)
- https://www.flickr.com/photos/130169972@N08/17180061387/in/dateposted-public/ (9)
- http://qr.net/JOJD (8)

- http://lin.io/Dd9p (8)
- http://twitter.com/WikiNut_Blue/status/633380345640054784/photo/1 (7)
- http://bit.do/7kwR (7)
- http://dlvr.it/Brz1yP (7)
- http://bit.ly/1Iq9Z8U (7)

The list of most popular retweets retrieved a long list of results, whose counted values were under 3. The top 5 of the most popular and most important retweets are:

- *Watch this one #news £ARYC ArrayIt http://t.co/2ehMWmaDk7 Check this out*
  Username: Yoannmv8544
  Times appeared: 18
- *#NEWS Date for the Victoria's Secret Fashion Show 2015 is November 10th in NYC! http://t.co/oEDgxmpvyx*
  Username: VSangelslove
  Times appeared: 13
- *Demi sera présente au NRJ Music Tour à Saint-Quentin le 5 septembre prochain. #news*
  Username: DEMI_SOOURCE
  Times appeared: 9
- *#Iran #News #Iraq: #UN rights office hails reform package #following countrywide protests http://t.co/aNJppNyB62 http://t.co/ZVjsqSOJlc*
  Username: Mojahedineng
  Times appeared: 7
- *GGreat poll #news from #Alabama!! @realDonaldTrump #1 • with a lead 2x's higher than #2 Jeb Bush. #Trump2016 http://t.co/enPcgxGT9M*
  Username: DanScavino
  Times appeared: 6

The Timeline of tweets reports the curves shown in Figures 5.3 and 5.4. According to the timeline, it is evident that the amount of information was the lowest on 5th of August (only 215 Tweets) but on the 15th of August there was almost 9-times more tweets (1924 tweets) with #news hashtag. On the 17th of August, the amount of data was a bit smaller but still very significant with 1601 tweets. When the view is switched to the

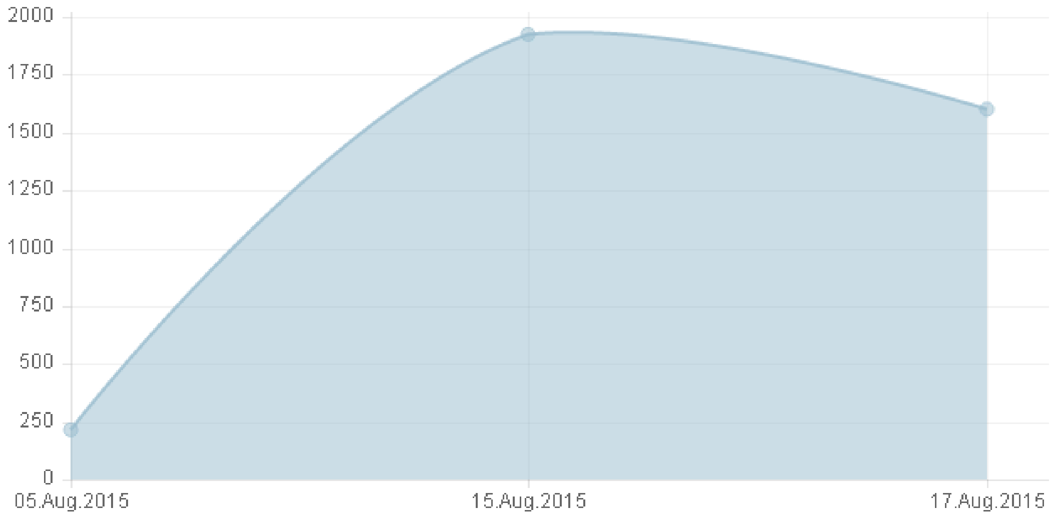


Figure 5.3.: Timeline of the results across days.

hourly-view, it can be determined that the largest peak of the tweets was collected between 18:00 and 19:00 o'clock on 15th of August and between 21:00 and 22:00 o'clock on 17th of August. It is obvious that the distribution of the tweets was the best in the evening rather than during the day. This fact is quite unusual since there is a general opinion that the news are being immediately distributed regardless of the used media. Essentially, the expectation of the news hashtag was that it should be equally distributed during the day and night, without apparent deviations. The fact that the most popular time for news broadcasting is between 18:00 and 22:00 brings us one steep closer to the possible answer to why are tweets are intensely presented during some time window and badly presented during some other time window.

The pie chart for the top words has retrieved some very interesting information about the most popular words and the number of their appearances. The results of the most popular words (including the number of appearances) are:

- with (240)
- will (144)

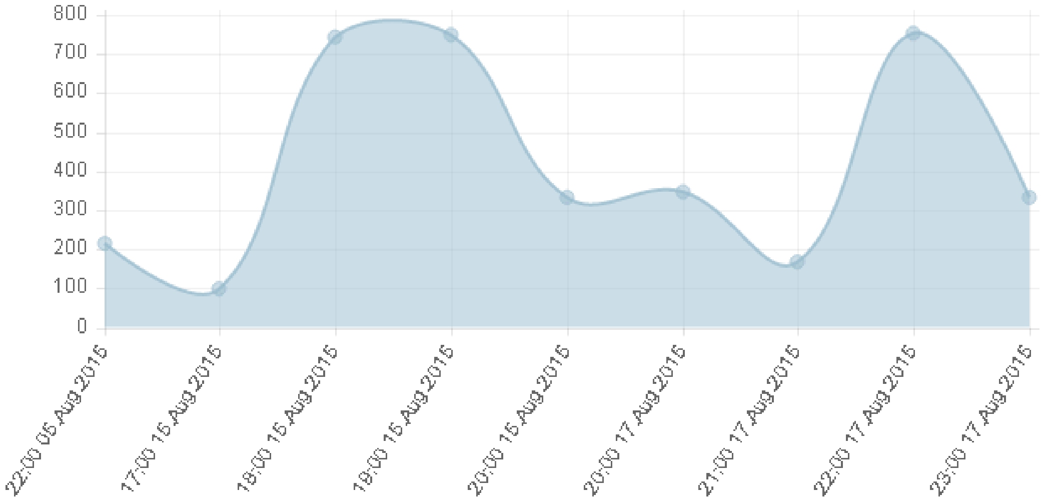


Figure 5.4.: Timeline of the results across hours.

- from (126)
- down (124)
- york (117)
- this (115)
- week (111)
- fashion (107)
- meet (104)
- news (101)
- teen (99)
- after (95)
- walk (90)
- syndrome (89)
- runway (89)
- video (85)
- suge (68)
- have (67)
- police (67)

The collection of the most popular words varies from nouns (week, fashion, police, teen), verbs (have, walk, meet) to prepositions (with).

The results of the mostly used software are as follows:

- twitterfeed (1011)
- IFTTT (854)
- dlvr.it (472)
- Twitter Web Client (172)
- Google (142)
- Twitter for Android (96)
- RSSGround (94)
- Twitter for iPhone (77)
- Mobile Web (M2) (72)
- FaceBOT-tw (39)
- Hootsuite (33)
- Facebook (30)
- RoundTeam (24)
- questoftruth21 (22)
- 2chfinder (22)

The majority of the elements (twitterfeed, IFTTT, dlvr.it, and so on.) are utility tools which are used for creation of the automatic tweets and posts with specific hashtags. Although they need the user's permission and are started manually, in a nutshell they are automatically created tweets which correspond to the rules for creation specified by the user. Sources like "Twitter for Android" or "Twitter for iPhone" can manually send tweets from various devices, such as Android cellphones or iPhones. Basically, it can be said that a great majority of tweets was sent by using services.

Figure 5.5 gives an overview of the most popular hashtags. The list of the exact results of the calculation is shown below. The list was restricted to the total of thirty elements according to the usability and readability of the results. As expected #tech hashtag occurred more often than any others hashtags which was expected since the term *tech* is related to technology which is a very popular news topic. The hashtag appeared exactly 198 times. The difference between the most popular hashtag (#tech) and the last one (#bbc which was found 33 times) is 166 appearances. The top ten hashtags could be assigned to general topics such as broadcasting, social network, politics, music and technology.

The list of Most Popular Hashtags:

- #tech (198)
- #rt (118)
- #midufinga (85)
- #iran (83)
- #newsdict (81)
- #music (79)
- #til (62)
- #pakistan (57)
- #setting4success (52)
- #straightouttacompton (51)
- #technology (49)
- #usa (48)
- #hiphop (46)
- #business (44)
- #politics (44)
- #lol (43)
- #world (43)
- #mobile (42)
- #socialmedia(42)
- #urdu (41)
- #entrepreneur(40)
- #funny (38)
- #dna (38)
- #health (36)
- #gadgets (36)
- #breakingnews (35)
- #breaking (34)
- #us (33)
- #uk (33)
- #bbc (32)

Figure 5.6 shows screenshots generated from the top twenty HTTP links. Due to the fast fluctuation of the information on the World Wide Web, it is possible that some of the HTTP links posted in the tweets are no longer available or they lead to articles which are outdated. The first screenshot in the first row in Figure 5.6 is an example of an outdated or expired link.

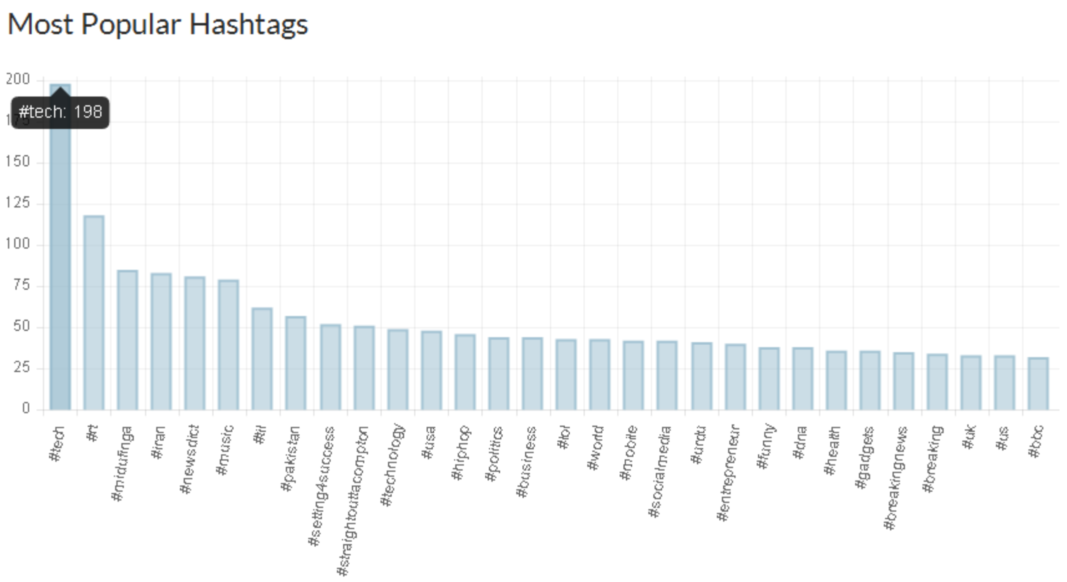


Figure 5.5.: Most popular hashtags for #news.

The Wikipedia results at the bottom of TwitterSuitcase are retrieving only four useful data objects. Those results are:

- *#tech* : The result is a partial result since the term ”tech” is commonly used as abbreviation for ”technology”. Hence, Wikipedia offers a list of possible articles related to the search item including the term description.
- *#rt* : Here, the results are basically the same as in the previous example but without the term description or definition. This result is also considered a partial result.
- *#news* : The result leads to the main Wikipedia article about the ”news” and such result is considered to be a full result according to 4.2.9.
- *#iran* : Also considered to be a full result since Wikipedia leads to the main article of the searched hashtag.

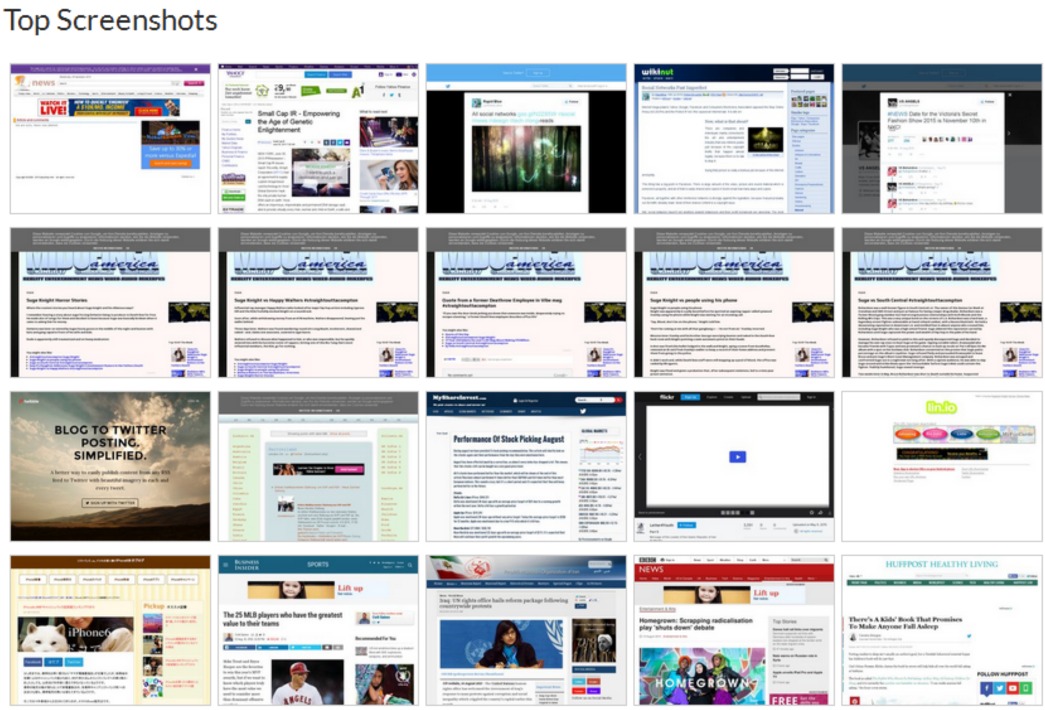


Figure 5.6.: Top Screenshots for #news.

5.1.3. Use Case Summary

The previous sections have provided an overview of the visualized Twitter data and a way to interpret them. It should give an answer to the research question outlined in chapter 1.1 Research Objectives: What kind or sort of information are we capable of providing during and after some event? The answer to this question is found when considering single parts of TwitterSuitcase (such as Top Users, Most Popular Hashtag, and so on.) since every mentioned part represents the analysis and visualization of the most and/or less important parts of the collected data. For the observer, the most important information within a single event are:

- Who wrote the most tweets?
- What are the most popular words or hashtags?
- What are the most popular RT's?
- When was the first and/or the last tweet sent?
- List of posted HTTP links?
- etc.

All of the above mentioned questions can be answered only by observing the data processed by TwitterSuitcase. Although this approach gives a good overview of the collected data about events, there is definitely some space for improvement, mostly within data analysis. The following chapter elaborates the future of Twitter data and its analysis.

5.2. Use Case #emoocs2014

5.2.1. Why #emoocs2014

EMOOC stands for European Massive Open Online Courses, a conference held each year on different locations across Europe, discussing E-Learning concepts and possibilities and presenting new ones. Since the conference is an event with a start and end date, it represents an ideal use case to test the functionality of TwitterSuitcase.

5.2.2. Using TwitterSuitcase with #emoocs2014

The tweets for this event were collected by TweetCollector. The next step, after data processing, is to create a TwitterSuitcase for the mentioned hashtag. The administration interface shown in Figure 5.1 gives the possibility of creating or deleting TwitterSuitcase data collections. The duration of the creation process for 4450 tweets was 260 seconds.

The collection of tweets included 4450 tweets for #emoocs2014 hashtags, which were published between 10th of February and 6th of August 2014.

From the word cloud, which represents all users according to the number of tweets they have published, it can be observed that the largest number of tweets came from the following accounts listed along with the number of posted tweets:

- moocf(185)
- Agora_Sup(141)
- fuscia_info(134)
- pabloachard(124)
- mooc24(120)
- tkoscielniak(103)
- bobreuter(85)
- OpenEduEU(84)
- yveszieba(81)
- redasadki(81)

These top ten users have published a total of 1138 or 25.6% of all tweets collected during this event. Figure 5.7 presents a word cloud and can be inspected for further use.

The list of most popular links, restricted to the top ten positions, has created the following results (including the number of occurrences):

- http://bit.ly/1la3yJX (32)
- http://twubs.com/eMOOCs2014 (28)
- http://www.emoocs2014.eu/sites/default/files/Proceedings-Moocs-Summit-2014.pdf (19)
- http://goo.gl/Ujaztf (16)

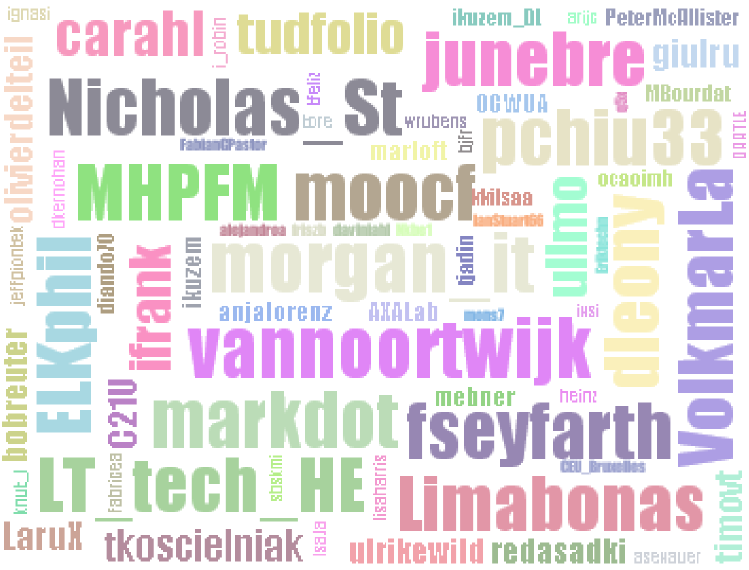
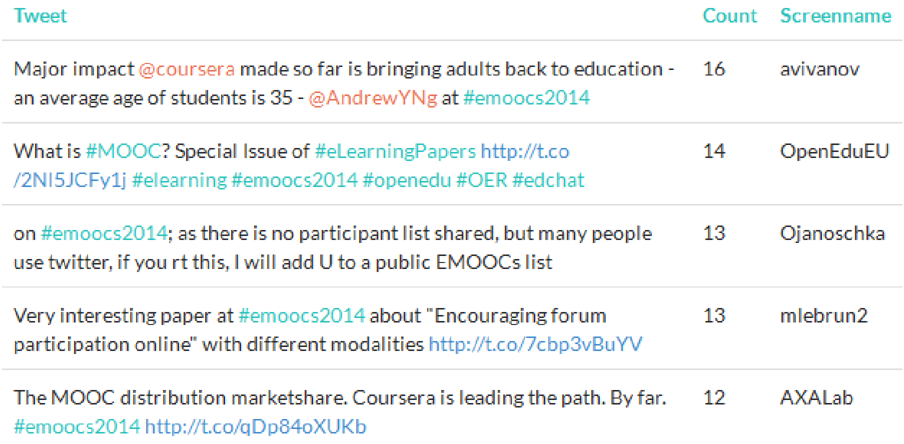


Figure 5.7.: Results for #emoocs2014 word cloud.

- `http://www.openuped.eu/` (15)
- `http://harvardx.harvard.edu/harvardx-insights/world-map-enrollment` (14)
- `http://rene.kizilcec.com/wp-content/uploads/2014/01/kizilcec_encour_emoocs2014.pdf` (14)
- `http://moocs.unige.ch/production/Growing_an_Edinburgh_MOOC.pdf` (14)
- `http://bit.ly/1cBV5wL` (14)
- `http://moby.to/rqxa95` (13)

The results of the most popular retweets are shown in Figure 5.8. The top five from the list of the most popular retweets are shown in the list below:

- *Major impact coursera made so far is bringing adults back to education - an average age of students is 35 - AndrewYNg at #emoocs2014*
  Username: avivanov
  Times appeared: 16
- *What is #MOOC? Special Issue of #eLearningPapers http://t.co/2Nl5JCFy1j #elearning #emoocs2014 #openedu #OER #edchat*



Tweet	Count	Screenname
Major impact @coursera made so far is bringing adults back to education - an average age of students is 35 - @AndrewYNg at #emoocs2014	16	avivanov
What is #MOOC? Special Issue of #eLearningPapers http://t.co/2NI5JCFy1j #elearning #emoocs2014 #openedu #OER #edchat	14	OpenEduEU
on #emoocs2014; as there is no participant list shared, but many people use twitter, if you rt this, I will add U to a public EMOOCs list	13	Ojanoschka
Very interesting paper at #emoocs2014 about "Encouraging forum participation online" with different modalities http://t.co/7cbp3vBuYV	13	mlebrun2
The MOOC distribution marketshare. Coursera is leading the path. By far. #emoocs2014 http://t.co/qDp84oXUKb	12	AXALab

Figure 5.8.: Most Popular Retweets for #emoocs2014 word cloud.

Username: OpenEduEU
Times appeared: 14

- *on #emoocs2014; as there is no participant list shared, but many people use twitter, if you rt this, I will add U to a public EMOOCs list*
  Username: Ojanoschka
  Times appeared: 13
- *Very interesting paper at #emoocs2014 about "Encouraging forum participation online" with different modalities http://t.co/7cbp3vBuYV*
  Username: mlebrun2
  Times appeared: 13
- *The MOOC distribution marketshare. Coursera is leading the path. By far. #emoocs2014 http://t.co/qDp84oXUKb*
  Username: AXALab
  Times appeared: 12

The timeline of tweets is shown in Figures 5.9 and 5.10. According to the timeline, the majority of tweets was collected at the very beginning of collection period, which was February 2014 with 4333 tweets or 97.3%, the remaining months from March until August 2014 have given only 117 tweets or 2.7% of all tweets. When we closely observe the tweets from February, we can estimate that 3080 tweets were collected during the first two days

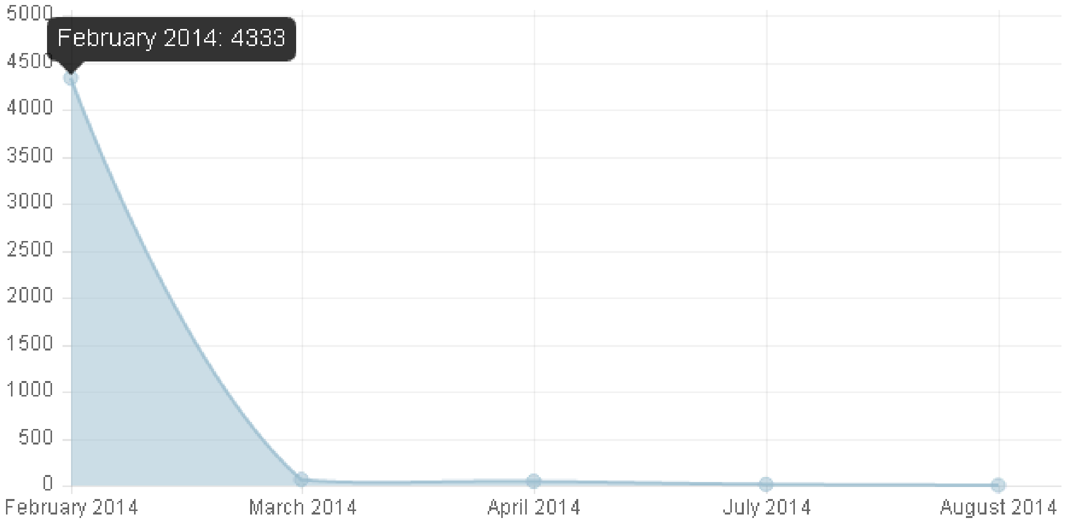


Figure 5.9.: Monthly view of timeline for #emoocs2014 hashtag.

which amounts to 69.2% of all tweets. The remaining 1253 tweets or 28.2.% were collected during the rest of February. Since the conference was held from 10th until 12th of February, the retrieved results were expected which means that almost 70% of all tweets considering a conference were posted during the conference.

Figure 5.11 presents the top twenty of the most popular words used in the tweets within an event. The best results were achieved by a Twitter "'RT"' shortcut with 2567 occurrences. Other words worth mentioning are moocs(802), mooc(639), learning(339) and openedueu(319). The rest of the words belong mostly to prepositions or articles.

Top used software or application (see Figure 5.12) shows that the majority of, precisely 1574 or 35.4%, were sent directly from Twitter using web service. The second largest group was represented through Apple devices (divided into various subgroups such as Twitter for iPhone, iPad or Mac) with a total of 1253 tweets or 28.5%. The third position is taken by a Twitter application called TweetDeck with 564 tweets or 12.7%, followed by Twitter for Android in the fourth place with 288 tweets or 6.5% of all tweets. The rest was split between various applications such as HootSuite, Mobile Web, Tweet Button,

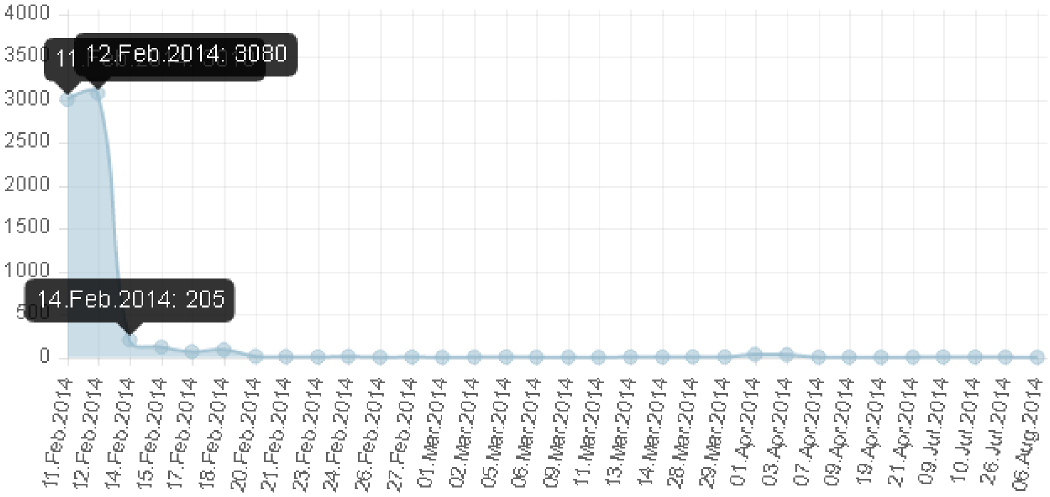


Figure 5.10.: Daily view of timeline for #emoocs2014 hashtag.

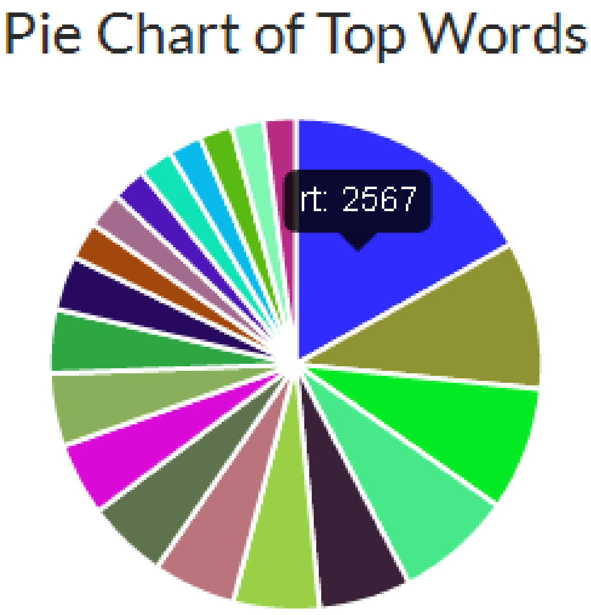


Figure 5.11.: Top Words for #emoocs2014 hashtag.

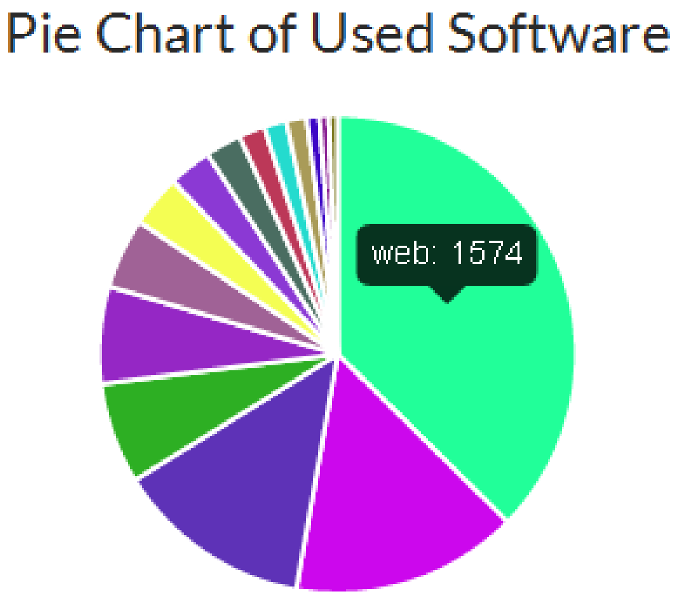


Figure 5.12.: Top Software for #emoocs2014 hashtag.

etc.

The listing below displays the top twenty hashtags used within the event and they are all undoubtedly related to the topics of conference. Figure 5.13 shows all values and the list shows hashtags including the number of occurrences:

- #mooc (308)
- #moocs (270)
- #elearning (72)
- #futurelearn (60)
- #vtecl (48)
- #elearningpapers (48)
- #bigdata (45)
- #heie (42)
- #edchat (42)
- #oer (39)
- #edtech (30)

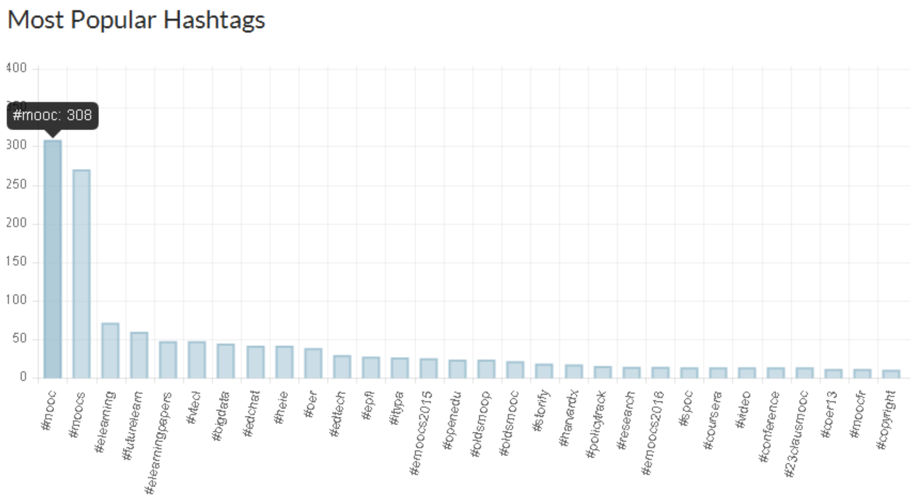
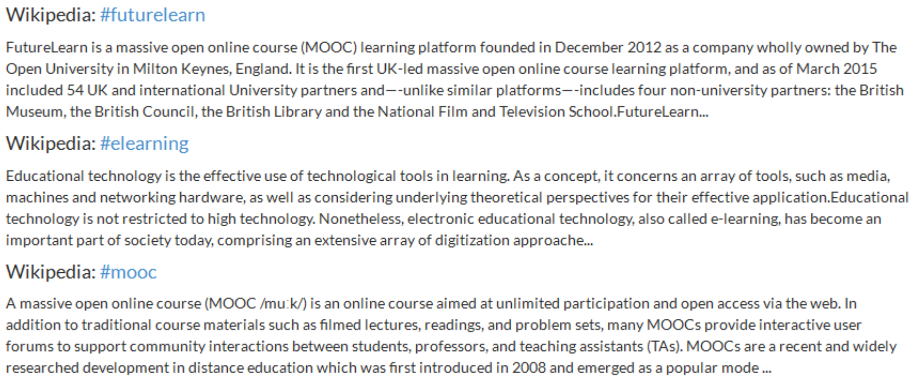


Figure 5.13.: Most Popular Hashtags for #emoocs2014.

- #epfl (28)
- #itypa (27)
- #emoocs2015 (26)
- #openedu (24)
- #oldsmoop (24)
- #oldsmooc (22)
- #storify (19)
- #harvardx (18)
- #policytrack (16)
- #emoocs2016 (15)
- #research (15)
- #coursera (14)
- #spoc (14)
- #video (14)
- #23clausmooc (14)
- #conference (14)
- #coer13 (12)
- #moocfr (12)
- #copyright (11)



Wikipedia: #futurelearn

FutureLearn is a massive open online course (MOOC) learning platform founded in December 2012 as a company wholly owned by The Open University in Milton Keynes, England. It is the first UK-led massive open online course learning platform, and as of March 2015 included 54 UK and international University partners and—-unlike similar platforms—-includes four non-university partners: the British Museum, the British Council, the British Library and the National Film and Television School.FutureLearn...

Wikipedia: #elearning

Educational technology is the effective use of technological tools in learning. As a concept, it concerns an array of tools, such as media, machines and networking hardware, as well as considering underlying theoretical perspectives for their effective application.Educational technology is not restricted to high technology. Nonetheless, electronic educational technology, also called e-learning, has become an important part of society today, comprising an extensive array of digitization approache...

Wikipedia: #mooc

A massive open online course (MOOC /muːk/) is an online course aimed at unlimited participation and open access via the web. In addition to traditional course materials such as filmed lectures, readings, and problem sets, many MOOCs provide interactive user forums to support community interactions between students, professors, and teaching assistants (TAs). MOOCs are a recent and widely researched development in distance education which was first introduced in 2008 and emerged as a popular mode ...

Figure 5.14.: Wikipedia results for #emoocs2014 hashtag.

The last two sections include the visualization of the most popular links in the form of snapshots and there are three full articles for #elearning, #mooc and #futurelearn hashtags on Wikipedia (see Figure 5.14).

6. Future of Twitter Data

The future of Twitter data is very encouraging since the popularity of social media and Twitter in particular is getting bigger by every day. The reasons for this popularity can vary from the accessibility of the service, over the velocity of the published information untill the importance of the published information.

The usage of Twitter for predicting some future events was proved by Asur and Huberman in their research published in "Predicting the Future with Social Media" [Asur and Huberman, 2010]. This research was based upon an assumption that the social media can be interpreted as a form of collective wisdom. The main objective of the researcher was to analyze this form of collective wisdom in order to predict real-world outcomes for a specific area of interest used by a common Twitter user. The focus was set on movies, since the topic of movies is considered very interesting among the majority of members of the social media community. In addition to that, the real-world outcomes in the form of a box-office results are easily reachable for comparison. The discussion between the users was observed and the assumption was made that the movies that are well talked about will also be well-watched and therefore achieve great box-office results. The observation included almost three million tweets and it has confirmed that there is a string correlation between the amount of attention a given topic has and its ranking in the future. The researchers have also established a model for predicting events within the social media.

The aforementioned article could be used as a crucial milestone in the usage and analysis of the Twitter data. Regardless of the fact that the Twitter data, with its 140 characters is very short and compact, it may contain very meaningful and important information, such as earthquake warnings or news updates. Since tweets are limited to the mentioned number of characters, there is no possibility for information growth. There is also a necessity to understand the meaning of the gathered information and checking the depth of that information through various analyzing algorithms is the only way to do that. The semantic analysis of the algorithms will certainly provide us with some new aspects of the information and give us an opportunity to understand the data by using some completely new approach. It is very difficult to predict the future of Twitter and its usage in the future. It is quite possible that Twitter will not be popular or existent in the next ten or twenty years. However, it has definitely established the concept of micro-blogging as one of the main trends which will also be exploited in the future.

7. Conclusion

The objective of this master's thesis was to determine what kind of information is available to a random user after an event is finished. Such analysis requires a wide research on several topics in this area of science.

Chapter 3: State Of The Art describes one of the ways to use Twitter data and explains the importance of such data collections. Furthermore, it provides an insight into the current scientific research in this area of expertise in terms of collecting, analyzing and visualizing tweets and their results. The scope of data collection includes information about users (such as location, current activities, etc.), lecture and scientific topics (such as conferences), news information (from various broadcaster like New York Times, CNN, etc.) and information about state of emergency (like earthquakes, period of civil unrest or even armed conflicts). The power of Twitter data could also be presented using the data to make predictions about some events in the future, for example outcomes of a box-office revenues for a specific movie (described in Chapter 6: Future of Twitter Data).

None of the analyzed and described tools could fully meet the research objective of this master's thesis and achieve satisfying results. Hence, a new tool was developed in order to meet the objectives. The developed tool called TwitterSuitcase was based on the collection of Twitter data previously created by TweetCollector (described in the first section of Chapter 4: TwitterSuitcase). The development of TwitterSuitcase mostly dealt with the visual representation of the collected tweets including partial analysis treatments. Once the data is analyzed, it will be presented in the form of word clouds, charts and various lists which are more comprehensible for a common user.

In order to test the functionality, a use case was applied in a separate chapter. The results which were obtained from the use case served as a basis for the discussion.

The implementation of the visualizing tool is successfully fulfilled but it is not complete. Since Twitter is a social media platform which is continuously growing, further implementation and extension of the tool will become and remain a consistent imperative. The usage of the APIs of TwitterSuitcase gives the possibility of integrating the tool within some other applications, websites or even services. Further development of TwitterSuitcase could be done both in the data analysis and the visualization of the data.

The influence of Twitter data is gaining in importance which makes it even more interesting for additional research. The mere accessibility of data is not enough. The data at hand must be processed properly in order to retrieve precise results for a specific scope. This thesis has shown that there are various software applications that could be used for data analysis and visualization but the majority of those applications could only be used within a specific scope of the research. The implementation fulfilled in this thesis can also be considered as a singular tool which can only be exploited for a specific scope. Due to the growth of Twitter, it is necessary to simultaneously develop various kinds of applications and tools to deal with the inflow of data which proved to be a very useful source of information in the present and in the future.

Bibliography

[Alexander Pak, 2010] Alexander Pak, P. P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *LREC*.

[Altmann, 2014] Altmann, T. (2014). Potential of twitter archives. Master's thesis, Graz University of Technology.

[Asur and Huberman, 2010] Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. *International Conference on Web Intelligence and Intelligent Agent Technology*.

[Boyd et al., 2010] Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. *Proceedings of the Forty-Third Hawai'i International Conference on System Sciences*, pages 1–10.

[Cha et al., 2010] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy. *Fourth International AAAI Conference on Weblogs and Social Media*.

[Chris Welch, 2015] Chris Welch, B. P. (2015). Twitter reaches 300 million active users, but the stock crashes after earnings leak early. `http://www.theverge.com/2015/4/28/8509855/twitter-earnings-q1-2015-leak-selerity`. Accessed: 2015-08-25.

[Ebner, 2013] Ebner, M. (2013). The influence of twitter on the academic environment. *Social Media and the New Academic Environment: Pedagogical Challenge*, page 293–307.

[Ebner and Maurer, 2008] Ebner, M. and Maurer, H. (2008). Can weblogs and microblogs change traditional scientific writing? *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, page 47–58.

[Ebner et al., 2008] Ebner, M., Reinhardt, W., Beham, G., and Costa, C. (2008). How people are using twitter during conferences. *Proceeding of 5. EduMedia conference*, pages 145–156.

[Hiruta et al., 2010] Hiruta, S., Yonezawa, T., Jurmu, M., and Tokuda, H. (2010). Detection, classification and visualization of place-triggered geotagged tweets. *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, page 956–963.

[Honeycutt and Herring, 2009] Honeycutt, C. and Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via twitter. *Proceedings of the Forty-Second Hawai'i International Conference on System Sciences*, pages 1–10.

[Huberman et al., 2008] Huberman, B. A., Romero, D. M., and Wu, F. (2008). Social networks that matter: Twitter under the microscope. *Preprint*.

[Jansen et al., 2009] Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, page 2169–2188.

[Java et al., 2007] Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65.

[Kaplan and Haenlein, 2011] Kaplan, A. M. and Haenlein, M. (2011). The early bird catches the news: Nine things you should know about microblogging. *Business Horizons, Vol. 54*, pages 89–103.

[Kumar et al., 2011] Kumar, S., Barbier, G., Abbasi, M. A., and Liu, H. (2011). Tweettracker an analysis tool for humanitarian and disaster relief. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

[Lavallee, 2007] Lavallee, A. (2007). Friends swap twitters, and frustration. http://www.wsj.com/articles/SB117373145818634482. Accessed: 2015-08-24.

[Neumann et al., 2010] Neumann, A., Barnickel, J., and Meyer, U. (2010). Security and privacy implications of url shortening services. -.

[Ng, 2013] Ng, J. Q. (2013). *Blocked on Weibo: What Gets Suppressed on China's Version of Twitter (And Why)*. The New Press.

[O'Reilly and Milstein, 2011] O'Reilly, T. and Milstein, S. (2011). *The Twitter Book*. O'Reilly Media.

[Petrovic et al., 2013] Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I., and Shrimpton, L. (2013). Can twitter replace newswire for breaking news? *Journal of the American society for information science and technology*.

[Richter et al., 2011] Richter, D., Riemer, K., and vom Brocke, J. (2011). Internet social networking, stand der forschung und konsequenzen für enterprise 2.0. *Wirtschaftsinformatik*, pages 105–113.

[Sakaki et al., 2010] Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. *Proceedings of the 19th international conference on World wide web*, pages 851–860.

[Twitter, 2015] Twitter (2015). Twitter turns six. https://blog.twitter.com/2012/twitter-turns-six. Accessed: 2015-08-25.

[Yousri Marzouki, 2015] Yousri Marzouki, O. O. (2015). Revolutionizing revolutions: Virtual collective consciousness and the arab spring. http://www.huffingtonpost.com/yousri-marzouki/revolutionizing-revolutio_b_1679181.html. Accessed: 2015-08-25.

[Zhao and Rosson, 2009] Zhao, D. and Rosson, M. B. (2009). How and why people twitter: The role that micro-blogging plays in informal communication at work. *Proceedings of the ACM 2009 international conference on Supporting group work*, page 243–252.

Appendix

Appendix A.

TwitterSuitcase Documentation

This chapter gives overview of various API calls for the tool.

A.1. TwitterSuitcase

TwitterSuitcase is based on a multiple API calls running across single PHP file located on the server. These API calls are described in the following section.

All results are retrieved in appropriate JSON data format.

A.1.1. Show all TweetCollector archives

Description

Retrieves the list of all twitter archives collected by TweetCollector and possible candidates for creating the TwitterSuitcase.

URL

`http://twitter.learninglab.tugraz.at/suitcase/api/data.php`

Parameters

- `suitcase=all` : Retrieves the list of all archives

Example of usage

`http://twitter.learninglab.tugraz.at/suitcase/api/data.php?archive=all`

A.1.2. Show all TwitterSuitcases

Description

Retrieves the list of existing TwitterSuitcase collections.

URL

`http://twitter.learninglab.tugraz.at/suitcase/api/data.php`

Parameters

- `suitcase=all` : Retrieves the list of all TwitterSuitcases

Example of usage

`http://twitter.learninglab.tugraz.at/suitcase/api/data.php?suitcase=all`

A.1.3. Create TwitterSuitcase

Description

Creates new TwitterSuitcase collection for a given archive ID.

URL

http://twitter.learninglab.tugraz.at/suitcase/api/data.php

Parameters

- `action` : Determines what kind of actions should be started. In this case ”create”.
- `id` : Archive ID within the database.
- `title` : Title of the newly created TwitterSuitcase.

Example of usage

The following example shows a possible API call:
http://twitter.learninglab.tugraz.at/suitcase/api/data.php?action=create&id=82&title=City%20of%20Graz%20TwitterEvent%202015

A.1.4. Delete TwitterSuitcase

Description

Deletes existing TwitterSuitcase data collection according to the given archive ID.

URL

http://twitter.learninglab.tugraz.at/suitcase/api/data.php

Parameters

- `action` : Determines the ”delete” action.
- `id` : Archive ID within the database.

Example of usage

The following example shows a possible API call:
http://twitter.learninglab.tugraz.at/suitcase/api/data.php?action=delete&id=82

A.1.5. Show TwitterSuitcase

Description

Retrieves all information of the TwitterSuitcase according to the archive ID. This represents the kernel of the TwitterSuitcase functionality due to the amount of retrieved data. Since the retrieved data is in JSON format it could be easily integrated into other applications or web services.

URL

http://twitter.learninglab.tugraz.at/suitcase/api/data.php

Parameters

- `action` : Determines the ”show” action.
- `id` : Archive ID within the database.

Example of usage

The following example shows a possible API call:
http://twitter.learninglab.tugraz.at/suitcase/api/data.php?action=show&id=82

A.1.6. Show all HTTP Links of TwitterSuitcase

Description

The result is a list of all collected HTTP links within the specific event processed by TwitterSuitcase.

URL

http://twitter.learninglab.tugraz.at/suitcase/api/data.php

Parameters

- `action` : Action ”show” leads to the HTTP details of the TwitterSuitcase.
- `id` : Archive ID within the TweetCollector’s database.
- `detail` : What kind of data should be shown in detail (HTTP).

Example of usage

The following example shows a possible API call:
http://twitter.learninglab.tugraz.at/suitcase/api/data.php?action=detail&id=82&detail=http

A.1.7. Show all tweets of TwitterSuitcase

Description

Retrieves the list of all collected tweets within the event.

URL

`http://twitter.learninglab.tugraz.at/suitcase/api/data.php`

Parameters

- `action` : Action "show" leads to the HTTP details of the TwitterSuitcase.
- `id` : Archive ID within the TweetCollector's database.
- `detail` : What kind of data should be shown in detail (tweets).

Example of usage

The following example shows a possible API call:
`http://twitter.learninglab.tugraz.at/suitcase/api/data.php?action=detail&id=82&detail=tweets`

A.1.8. Export HTTPs from TwitterSuitcase as CSV

Description

Creates new download file in Coma Separated Value (CSV) format with the list of all HTTP links including the number of times they have appeared.

URL

http://twitter.learninglab.tugraz.at/suitcase/detail.html?action=csv&id=79&detail=csv

Parameters

- `action` : Action "csv" leads to the export of all HTTP links of the TwitterSuitcase.
- `id` : Archive ID within the TweetCollector's database.
- `detail` : What kind of data should be shown in detail (csv).

Example of usage

The following example shows a possible API call:
http://twitter.learninglab.tugraz.at/suitcase/api/data.php?action=csv&id=79&detail=csv